NUREG/CR–6311
LA–12915-MS

Evaluating Prediction Uncertainty

Prepared by

Michael D. McKay
Statistics Group, Los Alamos National Laboratory
Los Alamos, NM 87545

Prepared for

U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Washington, DC 20555

February 9, 1995

ABSTRACT

The probability distribution of a model prediction is presented as a proper basis for evaluating the uncertainty in a model prediction that arises from uncertainty in input values. Determination of important model inputs and subsets of inputs is made through comparison of the prediction distribution with conditional prediction probability distributions. Replicated Latin hypercube sampling and variance ratios are used in estimation of the distributions and in construction of importance indicators. The assumption of a linear relation between model output and inputs is not necessary for the indicators to be effective. A sequential methodology which includes an independent validation step is applied in two analysis applications to select subsets of input variables which are the dominant causes of uncertainty in the model predictions. Comparison with results from methods which assume linearity shows how those methods may fail. Finally, suggestions for treating structural uncertainty for submodels are presented.

CONTENTS

FIGURES

EXECUTIVE SUMMARY

The importance of evaluating uncertainty in model predictions is well known to the Nuclear Regulatory Commission (NRC), as evidenced in the assessment of severe accident risks for five U.S. nuclear power plants in their NUREG–1150 report. Because of the need of NRC to quantify and understand uncertainty in model predictions, they have been and continue to be a driving force behind, for example, the use of simulation methods and Latin hypercube sampling (LHS) to estimate prediction probability distributions. Evaluation of the importance of inputs with respect to prediction uncertainty has been done, by and large, through regression-based methods including regression and correlation coefficients. Although regression-based methods have served well in providing importance indicators in many applications, they rely on linearity assumptions as the basis for their effectiveness. Breakdown of the assumptions can result in both failure to detect important inputs as well as false detections. Moreover, independent validation has been mostly overlooked as a means to confirm input selections and to provide estimates of the true effect of important inputs. These concerns constitute the motivation behind the present work to provide a sound theoretical basis together with effective methodologies for evaluating prediction uncertainty.

Vulnerable points in commonly used methodologies are addressed in three ways. First, the report shows that input importance can derive directly from the prediction probability distribution without reliance on specific assumptions such as linearity to relate model inputs and output. The report shows how regression-based methods can fail when such assumptions are invalid. Second, several types of variance ratios and sequential variable selection are shown to be reasonable and effective for identifying important inputs without dependence on linearity assumptions. Third, the value of validation to confirm input selections and to provide estimates of the true effect of important inputs is demonstrated in two analysis applications.

The methodology is shown to be effective in analysis applications for two very different models. In the first application—chosen because of the speed of model calculations—the flow of material in an ecosystem is described by a system of partial differential equations with 84 input parameters. Because the model is very fast running, subsets of inputs can be studied in detail. The second application involves the nuclear power station accident consequence analysis code called MACCS. Prediction uncertainty in three outputs arising from input uncertainty in 36 inputs is evaluated. This model runs very much slower than the first, and so abbreviated subset selection is employed. In both applications, sequential subset selection using variance ratios successfully identifies all important inputs. Thus, using the applications, the report illustrates analysis methods that can be applied to a wide variety of models.

The value of validation exercises to confirm input selection and quantify prediction uncertainty is illustrated in the analysis applications. Also illustrated through validation are the effects of important inputs which were undetected using partial rank correlation in the MACCS analysis.

The use of variance as an indicator of importance derives from a simple theoretical development of uncertainty using probability distributions. However, the idea that variance relates to importance is not new and, in fact, can be shown to underlie regression-based methods. Except for the cost of reliable estimation of variance in terms of the number of computer runs required—which can be significantly more than that required for estimation of partial correlation and regression coefficients—variance would be generally preferred over regression-based indicators in evaluation of prediction uncertainty. Fortunately, desktop computing makes variance estimation feasible in many applications.

Variance-based methods are made practical through development of a special LHS plan and heuristic procedures for selecting important inputs. The idea of the correlation ratio is extended to the partial correlation ratio, paralleling the partial correlation coefficient in linear models. The sequential procedures discussed in the analysis applications provide more complete pictures than are usually found of how different input combinations relate to prediction uncertainty.

Finally, the report addresses the topic of uncertainty in model predictions due to plausible alternative model structures—structural uncertainty. There is uncertainty in almost all model predictions because of approximate or incomplete treatment of the phenomenology of the process being modeled. For the most part, however, general treatment of structural uncertainty is virtually impossible owing to the conceptually large (possibly infinite) number of alternative models. The report presents a formal

basis for analysis and methods for analysis of structural uncertainty in submodel calculations.

In summary, practical methods for evaluating prediction uncertainty and a sound theoretical basis for them are presented. The methodology provides an effective description of the effects of input uncertainty that is more complete and defensible than those provided by commonly used regression-based techniques. Methods are illustrated in analysis applications.

ACKNOWLEDGMENTS

Several people at the Los Alamos National Laboratory deserve acknowledgment for contributing to this report. Masha Davidson provided many hours of technical support in making the MACCS computer runs for one of the analysis applications. Discussions with Lisa Moore and Richard Picard contributed valuable insights into sample design questions and into connotations of uncertainty and importance. Finally, Richard Beckman provided continual and indispensable counsel, ideas, and suggestions with regard to all aspects of the research.

At the NRC, James Glynn and Christiana Lui as technical monitors for the project provided helpful reminders of the realities of uncertainty analysis and were steady sources of encouragement.

1 INTRODUCTION

Evaluation of prediction uncertainty in a computer model means the estimation of the variability in model prediction due to uncertainty in input values and the determination of the contribution to the variability from dominant model inputs. When the mathematical form of a model makes analytical determinations impossible, simulation methods which include statistical estimation are used. Any of several sampling methods, like simple random sampling or Latin hypercube sampling (LHS), are suitable for estimation of prediction uncertainty. For the second part of evaluation, namely, identification of dominant inputs and their contribution to prediction uncertainty, regression methods are frequently employed. The most commonly used indicator statistics are correlation and partial correlation coefficients and regression coefficients, with and without the rank transformation. It is well known that statistics such as these derive their effectiveness from an assumed linear, or monotonic, relationship between model input variables and calculated prediction. Under that assumption, the variance of the model prediction is linear in input variances, and so the indicators properly attribute contributions of variability to different input variables. However, as the actual relation between inputs and prediction becomes less linear, the ability of regression indicators to function as intended diminishes to the point where they can fail completely to identify important inputs.

In order to development a cogent methodology for evaluating prediction uncertainty, this report begins with the probability distribution of a model prediction as a proper basis for evaluating the uncertainty. From that starting point, determination of important model inputs and subsets of inputs is seen to arise from comparison of the prediction distribution with conditional prediction probability distributions. From the many ways to compare probability distributions, a practical and intuitive one is through variances. The effectiveness of general variance-based indicators of importance does not depend on assumptions about the form of the relationship between inputs and prediction—in this sense, the indicators are nonparametric. It is not suggested that use of regression-based indicators should be discontinued; regression and correlation have served well and will continue to be necessary parts of analytical practices for long-running computer models. However, reliance on these indicators should be reduced in favor of nonparametric methods when the nonparametric methods are practical.

The methodology for uncertainty analysis of computer codes described in this report builds upon an earlier research project of the author sponsored by the United States Nuclear Regulatory Commission (NRC) and upon other extensive literature. The first research project began in 1975 when computer speed and costs severely limited the application of statistical analysis to reactor safety codes. Typically, the number of computer runs in an analysis would be no more than 20 and might require several weeks to complete at the nighttime computer charge rate. To accurately quantify the "error" or uncertainty in code calculations related to 15 to 30 input parameters seemed almost impossible. In this setting, LHS was developed by McKay, Conover, and Beckman (1979). It allowed successful assessment of sensitivity using partial rank correlation (McKay, Conover, and Whiteman, 1976) and, with the same set of computer runs, a reasonable measure of uncertainty ("error bands") with the tolerance interval (McKay and Bolstad, 1981).

The current project, begun in 1992, revisits the problem almost 12 years later, in a new computing setting where the desktop workstation can provide hundreds to thousands of runs a day. The effect of this new computing power is that what had been impossible—reasonable and effective description of model prediction uncertainty via probability distributions—is now a reality. Thus, current goals of analysis have expanded: they are higher, broader, and more optimistic than they were in the 1970s. Nevertheless, the fundamental objective remains the same: to quantify uncertainty in model predictions arising from uncertainty in input values and to identify principal contributors from among the model input variables and component submodels.

1.1 Overview

A mathematical model $m(\cdot)$ is a construction by which an output or prediction y is determined from a set of inputs x. Prediction uncertainty refers to the variability in prediction due to plausible alternative input values. The uncertainty about appropriate input values described by probability distributions propagates through the model to form a probability distribution for model prediction. The model prediction distribution provides the description of prediction uncertainty that is the object of investigation in this report.

Another source of uncertainty arises in almost all predictive or forecast models from their approximate or incomplete treatment of the phenomenology of the process being modeled. This source of uncertainty is termed structural or model uncertainty. A general characterization of structural uncertainty is much more difficult than one for input uncertainty. The notion of plausible alternative model structures is much larger than that of just alternative input values. It can include, for example, all continuous functions of an infinite number of input variables. Except for restricted classes of plausible alternative model structures (for example, when consideration is only among several competing models) the general treatment of structural uncertainty is virtually infeasible. Structural uncertainty is certainly of great importance. However, for the general case—the one usually encountered in reactor safety applications—practical methods for analysis have yet to be developed. Therefore, the prediction uncertainty discussed in this report, but for one exception, is that due to input uncertainty for an arbitrary but specified model structure. The exception is for submodel uncertainty. Under some circumstances, the effect of structural uncertainty of a submodel calculation might be evaluated relative to the effect of input uncertainty.

The application driving this work is the prediction of consequences from serious nuclear power reactor accidents. The input variables used in calculations in the computer codes describe initial conditions, release of radioactive material to the environment, transport of the material through the environment and the material's effects on people. The code—the model—developed from current understandings of physical processes through laws of physics and empirically derived associations, transforms input values into model predictions. At the focus of uncertainty analysis is the unknown difference between the model prediction and the outcome of an accident.

The difference between model prediction and truth is seldom known in the absolute sense outside of validation tests. Nevertheless, knowledge of the variability in prediction as input values change or different submodels are used is valuable to people who develop models and to those who use model predictions in the decision-making process. It is the goal of this report to present methods and procedures for uncertainty analysis which will accurately describe variability in model prediction and the contribution to that variability from various (subsets of) inputs.

1.2 Audience

The audience for this report is seen as consisting of two groups of people. Foremost, there are the technical people who build and test models and who must assess both adequacy and credibility of model prediction. Techniques of uncertainty analysis presented in this report can provide them with valid characterizations and descriptions of prediction uncertainty and input importance to use in their assessments. Moreover, the understanding of the extent of prediction uncertainty in model calculations is expected to contribute to their technical evaluations of models. It is assumed that model builders and the people who perform uncertainty analyses have a background in mathematics and statistics.

The other group of people in the audience for this report consists of decision-makers who use uncertainty analyses in their work. It is hoped that the material presented will illuminate the methods used in the uncertainty analyses so that both strengths and limitations can be better understood. Without doubt, the subject and methods of uncertainty analysis are mathematical. Nevertheless, the mathematical details of estimation of importance indicators can be passed over without loss to understanding.

From whatever background, however, the reader is assumed to be familiar with basic elements of probability theory, including the concepts of random variable, probability density function, and dependence of random variables.

1.3 Direction Taken in the Report

Various statistical procedures are used in evaluation of prediction uncertainty. Many of those used to identify important model inputs are borrowed, by and large, from regression analysis. Some others are based more generally on variance decomposition. As a background, the report presents a summary of these methods and references several good comparative studies. Interestingly, theoretical justification of many methods is strained, even when empirical studies indicate that they perform well. The need for a general and acceptable foundation for methods development and justification is apparent. To this end, the report examines modeling uncertainty in the abstract to develop a general notion of importance of inputs as being related to differences in probability distributions. Importance indicators formed from variance ratios then arise naturally from prediction variance, which is one manner through which the probability distributions might be compared. Statistical estimation of a variety of variance components used in importance indicators is then presented. With the theoretical development completed, procedures for performing uncertainty analyses are outlined and carried out on two sample applications. Finally, there is a short discussion of submodel uncertainty.

2 BACKGROUND DEVELOPMENT

Many statistical methods and practices can be collected under overlapping and sometimes indistinguishable umbrellas of model analyses variously called sensitivity analysis, sensitivity testing, error analysis, propagation of error, uncertainty analysis, and the like. From these are many that can be interpreted as foreshadowing or actually composing analysis of prediction uncertainty. This section reviews some of the methods from the point of view of analysis of prediction uncertainty. Apologetically, discussion of many applications of the wealth of methods from the past are omitted for the sake of brevity.

2.1 Two Perspectives for Performing Analyses

Two perspectives are used in model analysis (McKay, 1978 and 1988). One perspective focuses at points in the space of input values, like a nominal or base case, and is termed local relative to the input space. Historically, analysis from the local perspective has been called sensitivity analysis. The other perspective is from the space of output values or predictions. As such, its focus is not constrained a priori in the input space, and so it is termed global relative to the input space. It is from a global perspective that uncertainty analysis usually arises.

2.1.1 Local

From a local perspective, there is an input value x_0 of interest, for which knowledge of changes in the prediction y from small perturbations in inputs x about x_0 is desired. A common question in this situation concerns propagation of error, characterized by the derivatives of y with respect to the components of x. Objectives for study can be finding the direction, not necessarily parallel to a coordinate axis, in which y changes most rapidly or finding the change in y for an arbitrary direction. Issues like these lead to the concept of "critical" or "important" variables (or directions) as being ones which most account for change in y. For linear propagation of error, individual components of x are described as important or not. When the direction for change is arbitrary, meaning not necessarily along coordinate axes, subsets of the inputs which define direction, rather than individual inputs, become the issue. Typical of local analyses are one-at-a-time variational studies about the nominal input value.

2.1.2 Global

From a global perspective, interest lies in the event y exceeding (or not exceeding) specified values. Questions that arise in this case are concerned with associating particular inputs or segments of ranges of inputs with the event. Objectives of study might be related to controlling the event or to reducing its probability of occurrence in the real world by adjusting the values of some of the inputs. If costs are associated with the inputs, minimum cost solutions might be sought.

Clearly, both perspectives have a place in model analysis. In the local perspective, interest in x is restricted to a (small) neighborhood of a single point, and the derivative comes into play. In the global perspective, interest is in values of y, which might translate into a subset of, or possibly just a boundary in, the input space. In this case, the role of the derivative is less clear. What tends to blend the two perspectives is the use of the derivative to answer questions of a global nature. The practice is appropriate in small enough neighborhoods where the model is essentially linear, meaning that the derivative does not change substantially with x_0; or that, to first-order approximation, an "average" derivative is sufficient to characterize the model, again meaning that the model is essentially linear.

2.2 Prediction Uncertainty from a Global Perspective

Prediction uncertainty from a global perspective is different from uncertainty from a local perspective. Globally, the probability distribution of the prediction y contains all information about uncertainty without reference to input values. The distribution function can be estimated from a simple random sample of model runs. However, LHS is often a preferred alternative to simple random sampling (see McKay, Conover, and Beckman, 1979, and Stein, 1987). For both sampling methods, sampling error is a concern for small sample sizes.

Global uncertainty can arise from a local perspective by way of the relationship between y and x, often assumed linear and to hold over the entire input space. In this case, the characterization of uncertainty is usually through the variance of the prediction rather than through the whole probability distribution. How well the linearity assumption holds determines how well global uncertainty is characterized in this way.

2.3 Partitioning Prediction Uncertainty

Statements like “20% of the uncertainty in y is due to x_1” presupposes a quantitative measure and can be very misleading, depending on how well the probability distribution of y is summarized by the measure. An example of a more precise statement is “On average, the variance of y is 20% less when x_1 is fixed than when it is free; average is with respect to the distribution of x_1.” Variance is the natural but by no means unique candidate for a scalar measure of uncertainty.

Various methods address in one way or another the issue of partitioning or decomposing variance among inputs and subsets of inputs. Several studies compare and evaluate methods currently used in the analysis of computer models. Some of them are Saltelli, Andres, and Homma (1993), Saltelli and Homma (1992), Saltelli and Marivoet (1990), Iman and Helton (1988), and Downing, Gardner, and Hoffman (1985).

2.3.1 Linear Propagation of Error

When variance of y, $V[y]$, is the measure of uncertainty, the problem of partitioning uncertainty reduces to that of finding suitable decompositions for the variance of y. The simplest of these is the usual propagation-of-error method in which y is expressed as a Taylor series in the inputs x about some point x_0. To first-order approximation, the variance of y is expressed as a linear combination of the variances of the components of x by choosing x_0 to be μ_x, the mean value of x.

$$y(x) \simeq y(\mu_x) + \sum_i \frac{\partial y(\mu_x)}{\partial x_i}(x_i - \mu_{x_i})$$

$$V[y] \simeq \sum_i \left(\frac{\partial y(\mu_x)}{\partial x_i} \right)^2 V[x_i]$$

Derivatives might be determined numerically. Alternatively, Oblow (1978) and Oblow, Pin, and Wright (1986) use a technique whereby the capability of calculating derivatives is added into the (Fortran) model calculation using a precompiler called GRESS. When the derivatives of y are estimated by the coefficients from a linear regression of y on x, there seems to be a stronger assumption about the linear dependence of y on x. However, it is generally unknown whether the value of the actual derivative of y at $x = \mu_x$ or the value of an average slope is preferred in the variance approximation. In a technique that could be related to linear propagation of error, Wong and Rabitz (1991) look at the principal components of the partial derivative matrix.

Correlation coefficients have been used to indicate relative importance of the inputs. They are mentioned here because they are closely related to linear regression coefficients. In a similar way, rank-transformed values of y and x have been used for rank correlation and rank regression by McKay, Conover, and Whiteman (1976) and Iman, Helton, and Campbell (1981a, 1981b).

2.3.2 General Analytical Approximation

The natural extension of linear propagation of error, to add more terms in the Taylor series, makes it difficult to interpret variance decomposition component-wise for x. That is, the introduction of cross-product terms brings cross-moments into the variance approximation, which makes the approximation no longer separable with respect to the inputs. Nevertheless, higher-order terms in variance approximation may be necessary because of an obvious lack of fit from the linear approximation. The adequacy of the approximation to y might be used as a guide to the adequacy of the variance approximation.

Similarly, the linear approximation of y used in the regression can be generalized to an arbitrary analytical approximation from which, in theory, the variance of y can be derived either mathematically or through simulation. Alternatively, there is a method proposed by Sacks, Welch, Mitchell, and Wynn (1989) which looks at the model as a realization of a stochastic process. The difficulties in interpretation and assessing adequacy just mentioned for the higher-order Taylor series expansion apply here, too.

2.3.3 Sampling Methods

This final category of partitioning techniques relies on a sample (usually, some type of random sample) of values of y whose variability can be partitioned according to the inputs without an apparent assumed functional relation between y and x. In the category is a Fourier method of Cukier, Levine, and Shuler (1978). Their procedure samples values of each component of x in a periodic fashion, with different periods for each component. The variability (sum of squares) of the resulting values of y is written as a sum of terms corresponding to the different periods and thus associated with the different components. It is unclear how this relates to linear propagation of error, but it may be just another way to estimate the same quantities. The original Fourier method applies to continuous inputs; it is extended to binary variables by Pierce and Cukier (1981). Again, the relation to linear propagation of error is unclear. Another procedure suggested by Morris (1991) examines a probability distribution of the partial derivatives of the output arising from particular sampling designs.

Partition of variance in the multivariate analysis sense is becoming more important as an analytical tool. A most interesting partition of variance is presented by Cox (1982) from Baybutt and Kurth (1978) and is similar to a partition discussed by Karlin and Rinott (1982). It is given in Appendix A.1. Though not actually a sampling method, the elements of the decomposition are likely to be estimated from sampled data. The identity used involves the variances of conditional expectations of the output given subsets of the inputs. Iman and Hora (1990) use the expansion in its simplest form for a single input with an explicit polynomial approximation to the conditional mean. Saltelli, Andres, and Homma (1993) discuss a more general situation which relates to Krzykacz (1990) who uses the correlation ratio without an explicit form for the conditional mean. These ideas are discussed in detail in subsequent sections of this report. It is noted that, in general, it is not possible to construct a unique variance decomposition in which individual inputs are represented by single terms, one for each input.

2.4 Cautions

There are three important points about the methods just presented. First, uncertainty is only fully described by a probability distribution. Thus, while variance is often an effective characterization of the uncertainty, it can contain very limited information when the distribution is not symmetric with a long, heavy tail or when it is multimodal. Moreover, variance rarely characterizes the probability distribution uniquely. The second point, closely connected to the first, is that many methods identify individual inputs as important using variance under a linear approximation model. There are very few complete variance decompositions—for example, the Cox decomposition—which do not rely on some form of approximate relation between y and x. In particular, when linearity assumptions are invalid, ordinarily powerful methods based on them can break down. The final caution, which applies to all statistical procedures, concerns the part sampling variation plays in estimation. Different samples can produce very different estimates for the methods described. Thus, some type of independent validation of conclusions is prudent. The methods presented subsequently in this report address these cautions: differences in appropriate probability distributions are examined; methods apply to nonlinear models with only very weak assumptions; validation is employed for confirmation of conclusions.

2.5 Model Testing

Model testing is a term applied to a variety of procedures intended to evaluate and build credibility in a model's predictions. Although model testing logically precedes a final uncertainty study to evaluate prediction uncertainty, several aspects of it can be combined efficiently with a preliminary uncertainty analysis.

There are several main parts of model testing for any specific modeling application. It is expected that iteration among them will be necessary to achieve a reliable model. The parts are

- verification — determination of consistency between implementation of the model in a computer code and its conceptual or mathematical description
- calibration — determination of appropriate values of intrinsic model parameters that describe phenomenology
- shakedown testing — examination of model predictions for a wide range of input values
- validation — comparison between model predictions and experimental or observational data

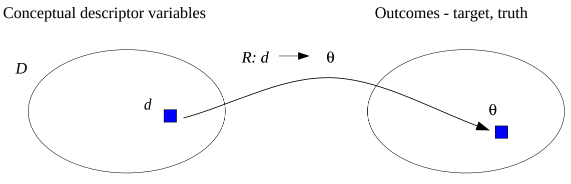
Model runs from a preliminary uncertainty study can be used in conjunction with the last two points of shakedown testing and validation. Simple visual displays of the data generated for an uncertainty study can provide a wealth of information because of the dispersion of input values in LHS. (See, for example, Ford, Moore, and McKay, 1979, and McKay, 1988.) In any event, a fully tested and validated model is necessary before prediction uncertainty due to input uncertainty can be evaluated sensibly.

3 MODELING UNCERTAINTY

Modeling uncertainty refers to the variability in model predictions due to plausible alternative input values (input uncertainty) and plausible alternative model structures (structural uncertainty). In this section, simple characterizations for input and structural uncertainty are proposed which allow formal description of uncertainty by probability distributions. It is pointed out that while structural and input uncertainty look very similar formally, structural uncertainty is fundamentally more difficult to evaluate in practice.

Modeling Uncertainty refers to the variability in model predictions due to plausible alternative input values (input uncertainty) or to plausible alternative model structures (structural uncertainty).

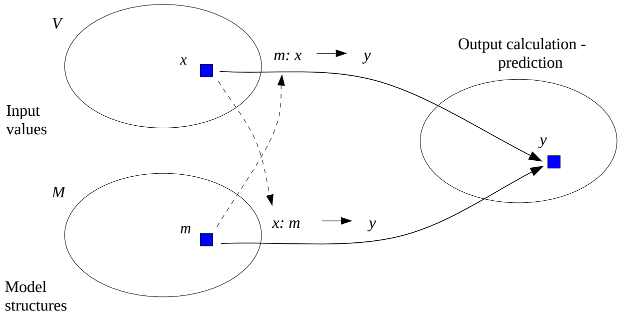
Following McKay (1993), models are mathematical abstractions, in the form of computer codes, used to predict outcomes of real events. One way to picture how outcomes arise in reality is depicted in Figure 3.1. Hypothetical descriptor variables d determine an outcome θ by their value and a rule R. The existence of descriptor variables is hypothetical; it is not critical to assume that a finite number of such variables actually exist and absolutely determine θ. The outcome θ might be a simple scalar or a vector, possibly of infinite dimension, discrete or continuous. It might be only partially observable or observable with error. The outcome θ might be a stochastic process governed by some components of d and specified by R. The rule R is unknown; formally, it maps descriptor variables d into outcomes θ.



$$\begin{aligned} d &: \text{ conceptual descriptor variables, } d \in D \\ R(\cdot) &: \text{ reality's rule or "law"} \\ \theta = R(d) &: \text{ target, outcome in reality} \end{aligned}$$

Figure 3.1 View of reality

The modeling process mirrors reality with input variables, structural form, and values of inputs by which the model output prediction is calculated. The modeling process is depicted in Figure 3.2.



$$\begin{aligned} x &: \text{ model inputs, } x \in V \\ m(\cdot) &: \text{ model structure, rule, algorithm, etc.} \\ y = m(x) &: \text{ model output calculation, prediction} \end{aligned}$$

Figure 3.2 Model of reality

Model predictions are often built upon idealizations and simplifications. They are calculated from presumed values for inputs with postulated relationships. In the upper part of Figure 3.2, a model $m(\cdot)$ from M is a map of the model input space V into the model prediction space. Model input space V and that of conceptual descriptors, D, need not coincide, as is the case, for example, when all relevant factors have not been identified. Specification of factors as model inputs is considered part of the model structure. In Figure 3.1, there is only one map, R, which is reality's unknown rule. Because of structural uncertainty, the possibility of alternative model structures is allowed in the space M of model structures. Looking at the situation from another angle in the lower part of Figure 3.2, a point x in the input space maps the space of models M into the prediction space. That is, for a specification of a situation through x, a range of possible predictions is spanned by varying $m(\cdot)$.

The term "model" is often used as if referring to a family of functions. Thus, a function might be thought of as a specific instance of a model.

Basic modeling elements are

$$\begin{aligned} x &: \text{ inputs, } x \in V \\ m(\cdot) &: \text{ structure, rule, algorithm, etc.} \\ y = m(x) &: \text{ output calculation, prediction} \\ \theta &: \text{ target of prediction} \end{aligned}$$

The model output y is a prediction of an unknown outcome θ and depends on both inputs x and model structure $m(\cdot)$. The prediction error for a simple, scalar prediction is the difference $y - \theta$. The two sources of prediction error are input values and model structure. These sources of error, or uncertainty, can be characterized formally in a probabilistic sense. Before doing so, however, an analogy is drawn from statistical analysis.

For fixed model structure $m(\cdot)$, prediction error $y - \theta$ follows from input uncertainty and comprises a component of precision (variance) and one of accuracy (bias) when y is treated as a random variable due to input uncertainty. The usual mean square error of prediction is

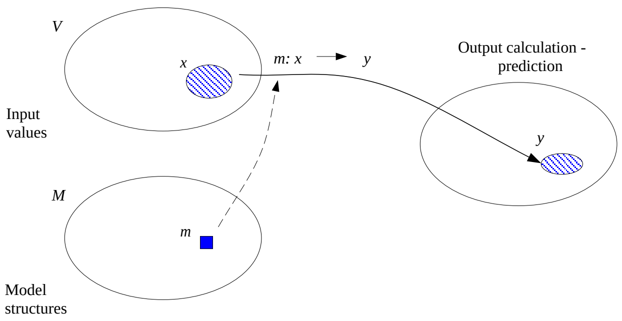
$$\begin{aligned} E\left[(y-\theta)^2\right] &= E\left[(y-\mu_y)^2\right] + (\mu_y - \theta)^2 \\ &= V[y] + (\mu_y - \theta)^2 \,, \end{aligned} \qquad (3\text{–}1)$$

where $\mu_y = E[y]$ is the mean value of y. The first term of the right-hand side of Equation 3–1 is the variance of y; it is a measure of prediction uncertainty called *prediction variance*. The second term on the right (bias squared) involves θ and, usually, cannot be evaluated. It measures the closeness of the average model prediction, μ_y, to θ and is a measure of accuracy. In any particular application, the expectation is with regard to input uncertainty for a fixed model.

There are really two fundamental sources of uncertainty in model prediction and the modeling process: (1) model structure—which identifies specific input variables and relationships among them—and (2) values of the inputs that specifically define modeled events. It may be important but impossible to consider both sources of uncertainty for a complete description of prediction uncertainty.

3.1 Input Uncertainty

Historically, model analysis has dealt almost exclusively with the input uncertainty component of modeling uncertainty. Input uncertainty refers to plausible alternative input values, as described in Figure 3.3. Corresponding prediction uncertainty, which depends on the model $m(\cdot)$, is indicated by the shaded area of variation in the prediction space associated with the shaded area of variation around in the input space V. No implications about the map, like continuity, are intended in the figure. By assumption, there is a probability function on the input space, represented by the density function f_x, which is mapped to f_y on the prediction space. The probability distribution f_x characterizes input uncertainty. The model $m(\cdot)$ can influence the choice of f_x. Therefore, the characterization of input uncertainty is the triple $(f_x, V, m(\cdot))$.



$m(\cdot)$: model structure
V: space of plausible input values, $x \in V$
f_x: probability function on V
f_y: probability function for y induced by $(f_x, V, m(\cdot))$

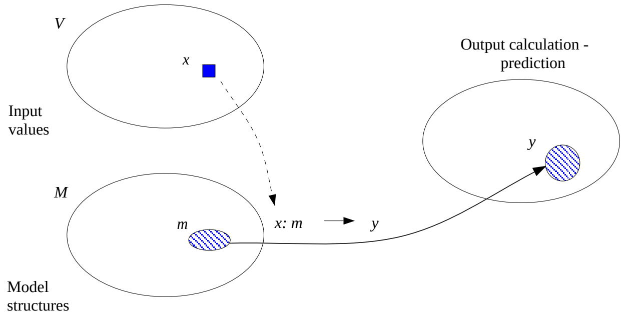
Figure 3.3 Characterization of prediction uncertainty from input uncertainty

Input Uncertainty refers to plausible alternative input values.

The evaluation of prediction uncertainty arising from input uncertainty concerns the determination or estimation of the range of variation of y, the estimation of the probability function f_y (or some of its moments), and some kind of determination of the "contribution" of various subsets of input variables to f_y. The evaluation of prediction uncertainty can be carried out using ordinary simulation methods.

3.2 Structural Uncertainty

The description of structural uncertainty parallels that of input uncertainty. Structural uncertainty refers to plausible alternative model structures, as described in Figure 3.4. The shaded area in M represents a “range” of alternative models which, for fixed input value x, induces the shaded neighborhood in prediction space. As before, the figure is not meant to suggest any particular properties, like continuity. Formally, there is a probability distribution on M represented by the density function g_m which induces a probability distribution on y indicated by the density g_y. This density represents the prediction uncertainty and depends on x. The characterization of structural uncertainty is the triple (g_m, M, x).



x: input value

M: space of plausible models, $m \in M$

g_m: probability function on M

g_y: probability function for y induced by (g_m, M, x)

Figure 3.4 Characterization of prediction uncertainty from structural uncertainty

Structural Uncertainty refers to plausible alternative model structures.

The evaluation of prediction uncertainty arising from structural uncertainty concerns the determination or estimation of the range, or space, of variability of y and the estimation of the probability function g_y. At this point the parallel between input uncertainty and structural uncertainty breaks down because of the difficulty of quantifying and sampling the space of models, M. A situation that is workable, however, is one where M consists of (finitely many) identified structures. This case is called one of competing models. It can be investigated in obvious ways, some of which are suggested by the discussion of submodel uncertainty in Section 10.

Another possibility involves representing models as realizations of stochastic processes. It has been used by Sacks, Welch, Mitchell, and Wynn (1989) and others for the purpose of designing computer experiments. Following their work, M would be a space of random functions, a superpopulation, whose parameters might be estimated from the model(s) at hand.

The formal definition of g_y as defining uncertainty in y is of little value without a basis for the probability distribution g_m on the space of models. Finding M and g_m constitutes the fundamental problem of structural uncertainty. For a decision-maker who must confront a situation involving structural uncertainty, Bayesian methods using subjective or degree-of-belief probability distributions (e.g., Apostolakis, 1990 and 1993) are available. However, application of Bayesian methods does not remove the difficulties of constructing fundamental probability distributions.

4 PREDICTION UNCERTAINTY

Prediction uncertainty describes the variability in the prediction y associated with input uncertainty. The important aspect of variability associated with structural uncertainty is not included because there are no general practical methods for treating it. In this section, fundamental ideas about uncertainty and importance are discussed with respect to the prediction probability distribution f_y. The concept of importance is related to differences among conditional probability distributions and prediction distribution. In Section 5, the variance of f_y is investigated as a summary for f_y to be used in evaluation of prediction uncertainty. Ideas of importance from conditional probability distributions carry over to the use of variance.

Prediction Uncertainty refers to the variability in y associated with input uncertainty and is characterized by the prediction probability distribution f_y. The model structure $m(\cdot)$ is assumed known and fixed, which is the usual case, and so the probability distribution is conditioned on $m(\cdot)$.

4.1 Prediction Probability Distribution

The probability distribution of the prediction y is represented by a probability density function f_y. (For discrete y, f_y is the usual probability function.) Conceptually, the density f_y derives from the input uncertainty triple

$$(f_x, V, m(\cdot)) ,$$

where

$$x \sim f_x \text{ for } x \in V$$
$$y = m(x) .$$

The density function f_x represents the probability distribution of the inputs x conditioned on the model $m(\cdot)$.

Theoretically, prediction uncertainty is completely described by f_y, the prediction probability distribution. Figure 4.1 depicts a density function f_y describing the probabilistic variability in y due to inputs x that vary over V according to f_x. Because f_y completely characterizes uncertainty in y, it cannot be discarded in final evaluations. However, in practical situations simple measures that are easy to use are needed. This is particularly true for when making model comparisons for different scenarios. Two widely used options are entropy and variance. Entropy, which plays a dominant role in information theory, is defined by

$$H = -E(\log(f_y)) .$$

Although possibilities for using entropy in uncertainty studies are very interesting, they are not yet well developed. A limited discussion is presented in Appendix A.2. The other option for summarizing uncertainty is the variance of f_y. This measure, dominant in the field of statistics, is investigated in Section 5.

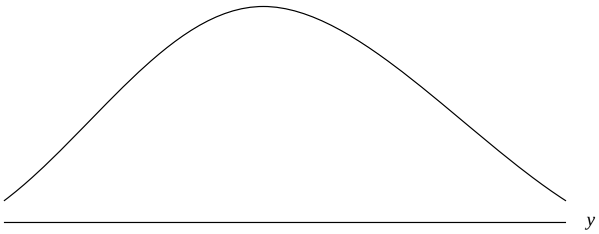


Figure 4.1 Prediction probability distribution f_y when all inputs vary

4.2 Importance for Prediction Uncertainty

Importance is a subjective and, hence, vague term. Operational definitions related to derivatives and regression coefficients, including partial correlation coefficients, may not be appropriate when dealing with prediction uncertainty. For studying prediction uncertainty, a convenient notion of importance relates to the “degree of dependence” between model prediction y and an input or subset of inputs. In the limiting case where y and an input are statistically independent, it is easy to understand that the input is not at all important: its value implies nothing about the value of y. At the other extreme, the value of an input could determine absolutely the value of y; that is, conditioned on the input, the value of y is fixed with probability 1. The input would be completely important. Somewhere in between these two limits is everything of practical interest.

For a general application model, the possibility that inputs individually are not particularly important needs to be considered. That is to say, no single input may have any particular impact on y, but collectively, a large enough subset of inputs will be important. Thus, the importance of a subset of inputs might depend more on the size of the subset than on its composition. Therefore, an important premise taken here is that importance of an input subset increases, or at least does not decrease, with the size of the subset.

The importance of input subsets includes importance of a single input. Let the inputs x be partitioned into disjoint subsets.

$$x = S_x \cup S_x^c$$

The importance of the subset S_x relates to the difference between f_y—the distribution of y when all inputs vary—and the family of conditional densities $\left\{f_{y|s_x}\right\}$ indexed on s_x and describing the variability of y when the subset S_x is fixed at different values s_x. As an example, Figure 4.2 suggests how S_x might reduce the variance in y for one of the densities in $\left\{f_{y|s_x}\right\}$.

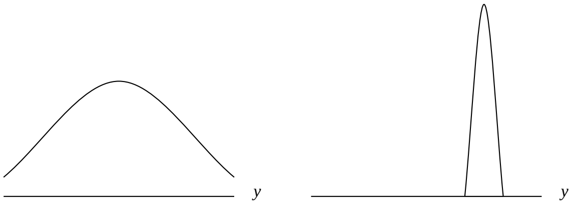


Figure 4.2 Unconditional f_y and conditional $f_{y|s_x}$ when some of the inputs are fixed

Figure 4.2 illustrates ideas of importance and local uncertainty: when the subset S_x of inputs is fixed at the value s_x, it is the remaining inputs in S_x^c which cause uncertainty in y. This local or conditional uncertainty is described by the conditional probability distribution $f_{y|s_x}$. The uncertainty is induced by the conditional input distribution $f_{s_x^c|s_x}$, which takes into account the possibility that the inputs may not be statistically independent. If S_x and S_x^c are independent, then

$$f_{s_x^c|s_x} = f_{s_x^c} \ .$$

Prediction uncertainty is described in terms of local uncertainty for any subset of inputs S_x. That is, the marginal (unconditional) density of y can be written as the average of conditional densities.

$$f_y(y) = \int f_{y|s_x}(y \mid s_x) f_{s_x}(s_x) ds_x \qquad (4\text{–}1)$$

The arguments of the density functions have been made explicit. The equation clearly shows the relation between f_y and the densities in the family $\left\{f_{y|s_x}\right\}$. Intuitively, that S_x is important means that the uncertainty in y changes with the values s_x. That is to say, that the densities in the family $\left\{f_{y|s_x}\right\}$ differ in some substantial way among themselves. On the other hand, that S_x is unimportant means that the fixed value s_x has small effect on y, which means that densities in the family $\left\{f_{y|s_x}\right\}$ are very similar among themselves. In the limiting case, S_x is *completely unimportant* when y and S_x are statistically independent. In this extreme case, the distribution of y conditioned on S_x is constant, namely,

$$f_{y|s_x} = f_y \text{ for all } s_x \ .$$

Thus, a base line for comparing the family of densities $\left\{f_{y|s_x}\right\}$ is the density f_y. Equation 4–1 shows that when the family $\left\{f_{y|s_x}\right\}$ are similar among themselves and all approach a constant function, that constant function is f_y. Likewise, when they are more dissimilar among themselves, they are dissimilar to f_y. The validation procedure discussed later in the report examines families of densities $\left\{f_{y|s_x}\right\}$ and $\left\{f_{y|s_x^c}\right\}$ and the density f_y.

Importance for Prediction Uncertainty refers to degree of statistical dependence between input and prediction and to differences within the family of conditional probability distributions $\left\{f_{y|s_x}\right\}$. The average value of the probability distributions in the family is the prediction probability distribution f_y.

4.3 Stochastic Variability

Prediction uncertainty relates to input uncertainty. However, there is another common type of variability, sometimes called stochastic uncertainty, which arises in connection with some modeling methods. For a stochastic model, the object of prediction is a random variable whose "stochastic variability" in nature is modeled as a random process. The behavior of the roll of dice or that of wind speed and direction at a weather station are examples of stochastic variability. Although reality and prediction can be summarized with histograms, means, and standard deviations, any particular output is random. Models of random processes like these are called stochastic process models or probabilistic models.

A complete and error-free specification of a stochastic (probabilistic) model can only provide predictions that are accurate in the statistical sense, for example, of being able to predict average behavior. For situations involving stochastic variability, it is convenient to view accuracy of prediction as referring to accuracy of the probability distribution from the model which describes the stochastic behavior of the actual outcome. In order to obtain an adequate estimate of the probability distribution of the stochastic outcome, a stochastic model must be sampled many times. When sampling error is significant with respect to input uncertainty, both types of variability need to be taken into account in an uncertainty analysis. In the remainder of this report, it is assumed that the probability distribution of stochastic outcome has been estimated, essentially, without error. Other treatments of stochastic models are left to further research efforts.

5 PREDICTION VARIANCE AND IMPORTANCE INDICATORS

In this section, prediction variance is presented as a simple measure of prediction uncertainty. The ideas are fundamental to the evaluation of prediction uncertainty and importance through prediction variance. In a development that parallels the general notions of prediction uncertainty importance previously discussed, variance is regarded as an attribute of and proxy for the prediction distribution f_y.

5.1 Prediction Variance

The mean value and variance of a probability distribution are fundamental attributes commonly used as a proxy for the full distribution, even though it is only in special cases, like that of the normal distribution, that the mean and variance uniquely identify the distribution. The mean and variance often contain enough information to suffice for analysis. Prediction mean and variance are given by

$$E(y) = \int y f_y(y) dy = \mu_y$$

and

$$\begin{aligned} V[y] &= E(y - \mu_y)^2 \\ &= \int (y - \mu_y)^2 f_y(y) dy \, . \end{aligned}$$

Investigation of importance with respect to prediction variance $V[y]$ follows.

5.2 Importance for Prediction Variance

For an arbitrary partition of the inputs x into disjoint subsets S_x and S_x^c, the (prediction) variance of y calculated from the left and right sides of Equation 4–1 produces the familiar result (see Parzen, 1962)

$$V[y] = V[E(y \mid S_x)] + E(V[y \mid S_x]) \, , \qquad (5\text{–}1)$$

where

$$V[E(y \mid S_x)] = \int \left(\mu_{y|s_x} - \mu_y\right)^2 f_{s_x}(s_x) ds_x$$

$$E(V[y \mid S_x]) = \int\int \left(y - \mu_{y|s_x}\right)^2 f_{y|s_x}(y) f_{s_x}(s_x) dy ds_x$$

and

$$\mu_{y|s_x} = E(y \mid S_x) = \int y f_{y|s_x}(y) dy \, .$$

5.2.1 Variance of the Conditional Expectation (VCE) of Prediction

The two terms on the right in Equation 5–1 have an interpretation as to the importance of the input subset S_x. The first of them is the variance of the conditional expectation of y, conditioned on S_x. It is denoted by VCE or

$$\text{VCE}(S_x) = V[E(y \mid S_x = s_x)] \, . \qquad (5\text{–}2)$$

The second term is an error or residual term written as

$$\text{Residual}(S_x^c; Sx) = E(V[y \mid S_x]) \, . \qquad (5\text{–}3)$$

Thus, the prediction variance is

$$V[y] = \text{ VCE}(S_x) + \text{Residual}(S_x^c; S_x) \, . \qquad (5\text{–}4)$$

The conditional expectation of y, also denoted by $\mu_{y|s_x}$, is a function of s_x, the conditioning value of the inputs in subset S_x. The VCE measures the variability in the conditional expected value of y as the inputs in S_x take on different values s_x. The residual term represents the variability in y not accounted for (explained by) the input subset S_x.

(In the notation for conditional expectation and variance, the notations " $\mid S_x$" and " $\mid S_x = s_x$" are used interchangeably to mean that the operation is conditioned on the subset S_x having an arbitrary but fixed value denoted by the lowercase symbol s_x.)

An informal argument that the VCE is a suitable measure for importance of the subset S_x follows. It looks at the constituent parts in Equation 5–4 to reveal the way in which they relate to S_x and the rest of the inputs S_x^c.

- The total variability in y when all of the inputs vary is measured by the prediction variance $V[y]$, the left side of Equation 5–4.
- When an arbitrary subset S_x is fixed at s_x, the expected prediction is given by conditional expected value of y, $E(y \mid S_x = s_x)$. It represents the prediction at s_x averaged over values of all of the other inputs S_x^c. The importance of S_x relates to how well S_x drives or controls y, that is, how well $E(y \mid S_x = s_x)$ mimics y. In particular, if the total variability in y is matched by the variability

in $E(y \mid S_x = s_x)$ as s_x varies, then S_x would be a very important input subset. That variability is measured by $V[E(y \mid S_x = s_x)]$, which is the VCE and the first term on the right in Equation 5–4.

- When S_x is fixed at the value s_x, the remaining or residual variability in y is due to all of the other inputs—S_x is fixed and S_x^c varies. The residual variability is the variability not controlled by S_x. It is measured by the conditional variance $V[y \mid S_x = s_x]$. The quantity is a measure of local variability at s_x due to S_x^c, and is averaged over s_x to yield its average $E(V[y \mid S_x = s_x])$, which is the second term on the right in Equation 5–4.

Equations 5–1 and 5–4 hold for continuous and discrete prediction variables y. They also hold when the subsets S_x and S_x^c are statistically dependent. When inputs are dependent, a large VCE for S_x might be more due to the conditional distribution of S_x^c changing with s_x than with the computational effect of S_x being fixed. This consideration points out the need to understand "importance" as it relates to the degree of statistical dependence of inputs and prediction.

In summary, the VCE is an intuitively appealing choice of an importance indicator. Equations 5–1 and 5–4 show that for any arbitrary subset of inputs, prediction variance can be written as the sum of a global component (VCE) and a local component (residual), from which the *correlation ratio,* discussed below, arises as a natural indicator of importance. No assumptions are made about the form of the relationship between y and x, as is the case for usual analysis of variance models and other regression models. However, convenient variance partitions which result from linear (approximation) models are not available. In Section 6, discussion of estimation procedures shows the relationship between the variance components used with importance indicators and traditional analysis of variance for random effects models.

5.2.2 Correlation Ratio

The constituents of the variance decomposition of Equation 5–4 are the VCE and the residual part due to the remaining inputs S_x^c. The magnitude of VCE relative to prediction variance in Equation 5–5 is called the correlation ratio by Kendall and Stuart (1979).

$$\begin{aligned} \eta^2 &= V[E(y \mid S_x)]/V[y] \\ &= \text{VCE}(S_x)/V[y] \end{aligned} \qquad (5\text{–}5)$$

They explain its use in describing nonlinear relationships as a parallel to that of the usual correlation coefficient ρ for linear relationships. A disincentive to its use in the past may have been the sample size required for adequate estimation of the VCE. The author had used an approximate estimator from LHS in early research efforts, but abandoned it due to imprecision in estimation. The same method is described by Krzykacz (1990) where the estimate is called the "empirical correlation ratio." Iman and Hora (1990) used $(V[E(y \mid x_j)])^{1/2}$ for single inputs $S_x = x_j$ in analysis of fault trees assuming a linear polynomial approximation for the conditional expectation of y. In Section 6, a new sampling plan for estimating correlation ratios is presented.

5.2.3 Partial VCE (PVCE)

The VCE and correlation ratio can be used as indication of importance of any specified subsets of inputs, including each input alone. It might be thought that to determine the composition of the important subset S_x, all that would be needed would be to assess each input separately using the VCE. Such a one-at-a-time approach, however, is not recommended because it is not necessarily an optimal or even good subset selection procedure. If, for p inputs in S_x, there were a unique partition of the VCE of the form

$$\text{VCE} = v_1 + v_2 + \cdots + v_p \ ,$$

where the v_i are nonnegative and correspond only to input number i, then selection of important input subsets would depend only on the relative sizes of the v_i. Unfortunately, there is no such unique partition, meaning that the relative importance of inputs is not well defined. This situation is similar to that in linear regression analysis where there is no unique partition of regression sums of squares—except when the x-values are orthogonal. Therefore, it is necessary to look further for a procedure for selection of subsets of important inputs.

In regression analysis, partial regression coefficients and partial sums of squares are used to select regression models. A similar procedure can be used for selecting important subsets of inputs in a model-free situation using variance components and the VCE. However, just as in regression, the order of variable selection will be seen to be material. The remainder of this section describes the procedure.

The heuristic sequential approach to assessing importance of inputs, finds "best" subsets of j inputs for $j = 1, 2, 3$

and so forth up to the smallest value for which all of the prediction variance is essentially accounted for. Proceeding sequentially, the subset S_x is augmented by one input variable, although it also can be (and, in practice, often is) augmented by a subset. The resulting VCE with an additional input is related to the initial VCE(S_x).

The VCE for the augmented subset

$$S_x^* = \{x^*, S_x\} ,$$

derived in Appendix A.3, is the sum of the VCE for S_x and an additional term for x^*. The new VCE is

$$\begin{aligned} \text{VCE}(S_x^*) &= V[E(y \mid S_x^*)] \\ &= \text{VCE}(S_x) + E(V[E(y \mid S_x^*) \mid S_x]) . \end{aligned} \quad (5\text{–}6)$$

The additional term is found by applying Equation 5–1 to the subset $\{x^*\}$ for each fixed value (at each site) of $S_x = s_x$. At each site, the prediction variance is conditioned on S_x and given by

$$\begin{aligned} V[y \mid S_x] = V[E(y \mid S_x^*) \mid S_x] \\ + E(V[y \mid S_x^*] \mid S_x) . \end{aligned} \quad (5\text{–}7)$$

Equation 5–7 parallels Equation 5–1 at each site $S_x = s_x$ and shows how importance of the additional input x^* beyond that of the subset of inputs S_x is evaluated locally with the conditional VCE as a function of s_x. A global measure is obtained by taking expectation (averaging) Equation 5–7 over S_x, from which the last term corresponding to x^* in Equation 5–6 can be derived. The expectation of Equation 5–7 with respect to S_x is

$$\begin{aligned} E(V[y \mid S_x]) = E(V[E(y \mid \{S_x, x^*\}) \mid S_x]) \\ +E(E(V[y \mid \{S_x, x^*\}] \mid S_x)) , \end{aligned} \quad (5\text{–}8)$$

which shows that Equation 5–1 can be written as

$$\begin{aligned} V[y] =& V[E(y \mid S_x)] \\ &+ E(V[E(y \mid \{x^*, S_x\}) \mid S_x]) \\ &+ E(E(V[y \mid \{x^*, S_x\}] \mid S_x)) , \end{aligned} \quad (5\text{–}9)$$

where the residual variance term is replaced by a term representing the additional variable (subset) x^* and a new residual term. The term for x^* is called the partial VCE or PVCE for x^* adjusted for S_x, and is given by

$$\text{PVCE}(x^*; S_x) = E(V[E(y \mid \{x^*, S_x\}) \mid S_x]) . \quad (5\text{–}10)$$

In terms of the VCE and the PVCE, the prediction variance in Equation 5–9 is

$$\begin{aligned} V[y] = \ & \text{VCE}(S_x) + \text{PVCE}(x^*; S_x) \\ & + \text{Residual}(S_x^{*c}; S_x^*) , \end{aligned} \quad (5\text{–}11)$$

where

$$\text{VCE}(\{x^*, S_x\}) = \text{VCE}(S_x) + \text{PVCE}(x^*; S_x) .$$

The representation forms the basis for the sequential construction of important subsets because the VCE for $\{x, S_x\}$ is as least as large as the VCE for S_x alone. That is, for two input subsets S and S^*,

$$S \subset S^* \rightarrow \text{VCE}(S) \leq \text{VCE}(S^*) .$$

This property is allows VCE to be used to construct subsets of important inputs sequentially, in a nested fashion.

5.2.4 Partial and Incremental Partial Correlation Ratios

The PVCE for x^* measures the amount of residual variance not explained by S_x that can be attributed to the additional input x^*. Relating the PVCE to the residual variance in Equations 5–1 and 5–3, yields the partial correlation ratio

$$\begin{aligned} \eta_p^2 &= E(V[E(y \mid \{S_x, x^*\}) \mid S_x])/E(V[y \mid S_x]) \\ &= \text{PVCE}(x^*; S_x)/\text{Residual}(S_x^c; S_x) . \end{aligned} \quad (5\text{–}12)$$

This ratio is an indicator of the (average) importance of x^* when S_x is fixed. If the PVCE is compared with the full prediction variance, the incremental partial correlation ratio is formed as

$$\begin{aligned} \eta_{\text{inc}}^2 &= E(V[E(y \mid \{S_x, x^*\}) \mid S_x])/V[y] \\ &= \text{PVCE}(x^*; S_x)/V[y] , \end{aligned} \quad (5\text{–}13)$$

which measures the importance of the x^* beyond (or adjusted for) that of the subset S_x.

The difference between VCE(x^*) and PVCE$(x^*; S_x)$ underlies the reason that the order is material in variable selection for identification of important inputs. This phenomenon means that the importance of an input by itself may be—and often is—different from its importance in concert with other inputs. Not only is it reasonable that this should be the case, but the phenomenon can be exploited when finding minimum-size or minimum-cost subsets to reduce prediction uncertainty.

5.2.5 Conditional Correlation Ratio

At each site $S_x = s_x$, prediction variance, VCE, and correlation can be computed for inputs *not* in the subset S_x. These quantities are called conditional, conditioned on $S_x = s_x$, and defined from Equations 5–7 and 5–5 in the obvious manner. Conditional VCE and conditional η^2 are local indicators of importance.

5.3 Conditional Moments and Model Testing

The moments in Equation 5–1 are integrals with respect to y which can be written also in the form

$$\begin{aligned} E(y) &= \mu_y \\ &= \int y f_y(y) dy \\ &= \int y(x) f_x(x) dx \end{aligned}$$

and

$$\begin{aligned} E(y \mid S_x = s_x) &= \mu_{y|s_x} \\ &= \int y f_{y|s_x}(y) dy \\ &= \int y(x) f_{s_x^c|s_x}(s_x^c) ds_x^c \end{aligned}$$

as integrals over the input space. The integrals suggest that estimators of the conditional mean

$$E(y \mid S_x = s_x)$$

and the conditional variance

$$V[y \mid S_x = s_x]$$

be used as diagnostic aids in model testing for evaluating input subsets along with their being components of importance indicators. Although plots of $E(y \mid S_x = s_x)$ can be very informative in revealing the effect of S_x, their interpretation should be weighted by the distribution f_{s_x} of S_x. In ordinary regression, for example, $E(y \mid S_x = s_x)$ is modeled as a function of s_x, and the importance of S_x is evaluated. Often, however, in ordinary regression analysis, the points are equally weighted. In the present situation, this corresponds to inputs having independent uniform probability distributions. The weighting is seen in the calculation of the (global) expected value of y as

$$E(y) = \int E(y(x) \mid S_x = s_x) f_{s_x}(s_x) ds_x \ , \qquad (5–14)$$

which is the weighted average of the (local) conditional expectations.

The conditional variance of y

$$V[y \mid s_x] = \int \left(y(x) - \mu_{y|s_x}(x)\right)^2 f_{s_x^c|s_x}(s_x^c) ds_x^c$$

can be used in a similar way as a diagnostic aid in model testing.

5.4 Beyond Regression Methods

Variance-based methods for assessing importance can be seen to be generalizations of regression-based methods by virtue of the treatment of the conditional expectation of y as a function of the conditioning variable x. Regression methods use an assumed form of the relationship, often linear. Variance methods operate without any such assumption. For example, in linear regression the conditional expectation of y is assumed to be a linear function of x, which can be written as

$$E(y \mid x) = x\beta$$

for a row vector x and column vector β. Under the linearity assumption, the VCE is given by

$$\text{VCE}(x) = V[E(y \mid x)] = \beta^t V[x] \beta \ ,$$

for which the unknown parameters β are estimated from the sample data. For variance-based methods, the VCE is estimated from sample data without regard to any specific relationship for the conditional expectation. In this sense, variance methods are model-free or nonparametric.

Estimation related to the VCE is directed to two components: estimation of the conditional expectation and estimation of its variance. With a linear regression approach, the entire estimation problem reduces to the ordinary regression analysis problem of estimation of the vector of parameters β. For a variance approach, the conditional expectation is an unspecified function of x. Therefore its estimation at each value of x and the subsequent estimation of its variance rely on sampling theory. Estimation for linear regression is well understood and requires a minimum number of computer runs, depending on the number of inputs and the complexity of the linear model. On the other hand, estimation for general variance methods requires many more computer runs. Therefore, variance approaches have been used in very limited situations in the past. Modern computing has opened the door to variance methods.

6 ESTIMATION FOR IMPORTANCE INDICATORS

This section discusses methods for estimating the variance components related to the VCE. Estimates are of the sums-of-squares variety from analysis of variance. Sampling plans are based on LHS. To more easily present fundamental ideas and more clearly present procedures used in analyses, estimation is discussed for three cases:

- $S_x = \{x_i\}$, individual inputs
- $S_x = \{x_i, S\}$, augmentation by individual inputs
- S_x an arbitrary subset of size $s > 1$

The first case occurs at the beginning of an uncertainty analysis when the objective is to assess the importance of each individual input with respect to prediction variance. A single sample of y-values based on LHS is used to estimate prediction variance and the VCE for each input variable. The second case is encountered when sequentially constructing and assessing importance of subsets S_x. The case examines a method for estimating the PVCE, the increment to the VCE from the addition of inputs x_i to a previously selected subset S. As in the first case, the same sample of y-values is used for each input variable. The final case arises in evaluation of an arbitrary subset of inputs. It is a simple extension of the first case with the subset treated as a single input variable.

The estimators are sums of squares and arise in a natural manner from familiar analysis of variance formulas, as illustrated in Appendices A.4, A.5, and A.6.. Their properties come from simple random sampling and carry over, through approximation, to LHS. Estimators corresponding to different subsets S_x are not required to be independent, and it is unlikely that they are. It is emphasized that there is no assumption that $y = m(x)$ is a linear function of the inputs x.

6.1 Basic Sampling Plan: Replicated LHS

Variance components are estimated from a design called a replicated LHS (rLHS). Each replicate in an rLHS corresponds to independent randomizations of the set of values of each input. An rLHS is not a replicated LHS plan in the usual sense that replicates are independent and identically distributed samples. However, even in an LHS, the individual sample values are not independent and identically distributed samples.

The properties of an rLHS based on an LHS of size n are conditioned on the n particular values of each input. In an LHS, values are sampled from intervals, a procedure that allows certain estimators to be unbiased. A disadvantage of the procedure is that some samples will contain extreme values (in the tails of the distribution) that cause sample estimates to be unusually far from the true value. As a compromise, probability midpoint or median values from each interval are used instead of sampled values. This change is equivalent to changing the input space from continuous to discrete. From a practical point of view, the change is not likely to be material. Theoretically, for a large enough number of intervals and suitably smooth models $m(\cdot)$, the use of probability midpoints is also immaterial and might even produce better estimators in a mean-square-error sense.

6.1.1 Base Case Sample

An LHS of size n for I inputs is denoted by the matrix

$$\boldsymbol{D}_0 = [X_1, X_2, \cdots, X_I]$$

of dimension n rows × I columns. Each column vector $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})^t$ contains n values x_{ij} sampled from equal-probability intervals and randomized as to position in the vector. Although not crucial to the design, probability midpoints of the intervals rather than sampled values are used.

An rLHS-n is r replicates of the LHS obtained as independent permutations of all of the columns of $\boldsymbol{D}_0$. Replicate k and the full design $\boldsymbol{D}$ are denoted by

$$\boldsymbol{D}_k = \left[\widetilde{X}_{1,k}, \widetilde{X}_{2,k}, \cdots, \widetilde{X}_{I,k}\right],\ k = 1, 2, \cdots, r$$

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_r \end{bmatrix}, \qquad (6\text{–}1)$$

where $\widetilde{X}_{i,k}$ is an independent permutation of the rows of X_i. The full design matrix $\boldsymbol{D}$ is an $(r{\times}n)$ row × I column matrix. The construction points out that the same n values for each input appear in each of the r replicate design matrices and that the replicate design matrices differ in the input combinations designated by the rows of the matrices.

6.1.2 Dependent Inputs

The situation is certainly simpler when inputs are independent. Because of this, inputs are sometimes treated as independent when it is more appropriate to treat them as dependent. Alternatively, dependence among inputs can be approximated by inducing sample correlation structure through a permutation process, such as described by Iman and Conover (1982), or by a procedure due to Stein (1987). Finally, when a proper treatment is required—as when the ranges of inputs depend on each other—sample values are selected according to their joint probability distribution. This action can be accomplished in two ways, depending on the specific situation. If, for example, x_1 and x_2 are not independent and have joint density function $f_{12}(x_1, x_2) = f_1(x_1) \cdot f_{2|1}(x_2 \mid x_1)$, then one method is to sample directly from f_{12}, and the other method is to sample first from the marginal density f_1 and then from the conditional density $f_{2|1}$. For best estimation of f_y, the proper distribution of x_1 and x_2 should be used. However, for determining important inputs, approximate sampling methods can be used during screening followed by a proper sampling method for validation.

6.2 Estimation for Individual Inputs

Estimation for assessing importance of individual inputs is a fundamental component of uncertainty analysis. This section discusses sample design, formulation of estimators, and critical values. The general principles used readily extend to the augmentation and arbitrary subset cases presented subsequently.

6.2.1 Sample Design

Estimation of variance components for all of the individual inputs can be accomplished using a single rLHS of size n with r replicates. The full sample requires $N = n \times r$ model runs and predictions y. (As mentioned in the introduction to this section, it is neither required nor likely that the estimates are independent.) The design matrix $\boldsymbol{D}$ for r replicates in an rLHS-n is given in Equation 6–1. The associated model predictions y are $\{y_{jk}\}$ for $j = 1, \cdots, n$ and $k = 1, \cdots, r$.

6.2.2 Prediction Variance Estimate

Each of the r replicates of an LHS yields an estimate of the variance of the prediction y from each replicate sample (total) sum of squares as

$$\widehat{V}_k[y] = \frac{1}{n} \sum_{j=1}^{n} \left(y_{j\,k} - \overline{y}_{\cdot\,k}\right)^2$$
$$\overline{y}_{\cdot\,k} = \frac{1}{n} \sum_{j=1}^{n} y_{j\,k}\ .$$

The n-divided sum of squares is preferred to the $(n-1)$-divided one with LHS. For simple random sampling, the $(n-1)$-divided sum produces unbiased estimators. The familiar analysis of variance relationship for sums of squares between replicates and within replicates is

$$\sum_{k=1}^{r} \sum_{j=1}^{n} \left(y_{j\,k} - \overline{y}\right)^2 = \sum_{k=1}^{r} \sum_{j=1}^{n} \left(\overline{y}_{\cdot\,k} - \overline{y}\right)^2 + \sum_{k=1}^{r} \sum_{j=1}^{n} \left(y_{j\,k} - \overline{y}_{\cdot\,k}\right)^2\ ,$$

where

$$\overline{y} = \frac{1}{r} \sum_{k=1}^{r} \overline{y}_{\cdot\,k}\ .$$

It shows that the pooled variance estimator is

$$\begin{aligned} \widehat{V}_p[y] &= \frac{1}{r} \sum_{k=1}^{r} \widehat{V}_k[y] \\ &= \frac{1}{nr} \sum_{k=1}^{r} \sum_{j=1}^{n} \left(y_{j\,k} - \overline{y}_{\cdot\,k}\right)^2 \\ &= \frac{1}{nr} \sum_{k=1}^{r} \sum_{j=1}^{n} \left(y_{j\,k} - \overline{y}\right)^2 - \frac{1}{r} \sum_{k=1}^{r} \left(\overline{y}_{\cdot\,k} - \overline{y}\right)^2\ . \end{aligned}$$

The pooled estimator is only close to unbiased—note the divisors in the sums of squares—for simple random samples of size nr. However, it is even less biased for LHS. For LHS, the last term involving the replicate means is expected to be small, so the estimator used for variance of y is the simpler form

$$\widehat{V}[y] = \frac{1}{nr} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{j\,k} - \overline{y}\right)^2\ . \qquad (6\text{–}2)$$

6.2.3 VCE Estimate

The VCE for x_i given by Equation 5–2 is

$$\mathrm{VCE}(x_i) = V[E(y \mid x_i)] \; .$$

It is estimated separately for each input. Without loss of generality, the predictions y are assumed to be labeled so that $\{y_{jk},\ k = 1, \cdots, r\}$ corresponds to x_{ij}. Estimation of the VCE is viewed in two parts: one concerning the expected value (E) and the other concerning the variance (V). The estimator in the expectation part is the sample average of y-values for which $x_i = x_{ij}$:

$$\overline{y}_j = \frac{1}{r} \sum_{k=1}^{r} y_{jk} \; .$$

For simple random sampling and, approximately, for LHS,

$$E\left(\overline{y}_j\right) = E(y \mid x_{ij})$$

as desired. The expected value part of the estimation is based on r values y_{jk}.

The complete construction in the variance part begins with the sum of squares whose expectation, from Appendix A.4, is

$$E\left(\frac{1}{n} \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 \right) \simeq V[E(y \mid x_i)] + \frac{1}{r} E(V[y \mid x_i]) \; .$$

Therefore, the VCE for x_i can be estimated with

$$\widehat{\mathrm{VCE}}(x_i) = \frac{1}{n} \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 - \frac{1}{nr^2} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{jk} - \overline{y}_j\right)^2 . \qquad (6\text{–}3)$$

The variance part of the estimation is based on n values of the mean $\overline{y}_j$.

6.2.4 Correlation Ratio Estimate

The correlation ratio is estimated by a ratio of estimators in

$$\widehat{\eta}^2 = R_a^2 = \widehat{\mathrm{VCE}}(x_i) / \widehat{V}[y] \; , \qquad (6\text{–}4)$$

which, in terms of Equations 6–2 and 6–3, is

$$\begin{aligned} R_a^2(x_i) &= \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 - \frac{1}{nr^2} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{ij} - \overline{y}_j\right)^2 \right\} \\ &\qquad / \left\{ \frac{1}{nr} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{ij} - \overline{y}\right)^2 \right\} \\ &= \left\{ r \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 - \frac{1}{r} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{ij} - \overline{y}_j\right)^2 \right\} \\ &\qquad / \left\{ \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{ij} - \overline{y}\right)^2 \right\} . \qquad (6\text{–}5) \end{aligned}$$

These equations are made up of the sums of squares from a one-way analysis of variance. How they relate the VCE and residual variance components with analysis of variance is shown in Appendix A.5.

Importantly, while analysis of variance ordinarily applies with a linear model, the estimators of prediction variance and VCE used in R_a^2 do not depend upon any such model. Estimates of the correlation ratio and partial correlation ratio show how analysis of variance formulas relate to estimation of variances used for importance indicators. In fact, the quantity

$$R^2 = \left\{ r \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 \right\} \Big/ \left\{ \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{ij} - \overline{y}\right)^2 \right\} \qquad (6\text{–}6)$$

from a linear (analysis of variance) model is related to R_a^2 through

$$R_a^2 = R^2 - \frac{1}{r}\left(1 - R^2\right) . \qquad (6\text{–}7)$$

Derivation of critical values for R^2 and R_a^2 follows.

6.2.5 Critical Values

Critical values for R^2 and R_a^2 are derived under the null hypothesis for a random-effects model that y_{ij} are $n \times r$ independent and identically distributed normal random variables partitioned at random into n groups of size r. The null hypothesis implies that the labeling y_{jk} of y-values according to the values x_{ij} constitutes a random partition—that y is independent of x_i. The additional assumption of (approximate) normality is common and needed for R^2 to have a beta distribution. The beta distribution is related to the F distribution through the transformation

$$R^2 = \frac{1}{1 + (\upsilon_2/\upsilon_1)\mathrm{F}(\upsilon_2, \upsilon_1)} \,, \tag{6–8}$$

where

$$\begin{aligned} \upsilon_1 &= n - 1 \\ \upsilon_2 &= n \cdot (r - 1) \,. \end{aligned}$$

The expected value of R^2 under the null hypothesis is

$$\begin{aligned} E\big(R^2\big) &= \frac{n-1}{n \cdot r - 1} \\ &\simeq \frac{1}{r} \,, \end{aligned} \tag{6–9}$$

which shows how large, and small, R^2 is expected to be as a function of the number of replicates r. Equation 6–7 is used to transform values from R^2 to R_a^2.

6.2.6 Notes on Approximations

The expectation results are derived from a simple random sample of observations y_{ij}. Additionally, the beta distribution of R^2 derives from the y_{ij} having a normal distribution. Therefore, the results are approximate for an LHS and for y_{ij} which are rank transformed. Nevertheless, the critical values and mean value for R^2 provide convenient practical guidelines for analysis.

6.3 Estimation for Augmentation by Individual Inputs

Augmentation describes a situation very much like the one treating individual inputs, except that importance of inputs is assessed over and above the importance of a previously chosen subset of inputs. For a previously chosen subset S_x, the objective through augmentation is to estimate the VCE for subsets $\{x, S_x\}$ which include an additional input. The strategy is to use the relationship, from Equation 5–11,

$$\text{VCE}(\{x_i, S_x\}) = \text{VCE}(S_x) + \text{PVCE}(x_i; S_x) \,. \tag{6–10}$$

The PVCE is estimated as the average of the (conditional) VCE estimates over a sample of values (sites) of S_x.

6.3.1 Sample Design

The subset S_x is a subset of inputs. Let

$$v = \{v_1, v_2, \cdots, v_s\}$$

denote s sites for S_x, meaning s vectors that give the values for the inputs in S_x. The design matrix at site t is denoted by

$$\boldsymbol{D}_t = v_t \otimes \boldsymbol{D}_{(S_x)} \,, \tag{6–11}$$

meaning that the columns of $\boldsymbol{D}$ in Equation 6–1 corresponding to the inputs in S_x are replaced by the fixed values in v_t. The values of the inputs S_x are constant and equal to the values v_t at site t. Finally, the sites are selected by LHS.

It is allowed that only one site of S_x be sampled. This situation is equivalent to setting the inputs in S_x to their median values in the base case sample and doing the analyses for individual inputs.

6.3.2 Conditional VCE and PVCE Estimates

At each site t, the base case design matrix is used for all inputs except those in S_x, which have fixed values at each site. The VCEs at site t, calculated with the predictions $\{y_{tjk}\}$, are conditioned on the value v_t of S_x and called *conditional* VCEs. When averaged over all sites t, they become the PVCE for each input. Applying Equation 6–3 at each site gives the conditional VCE estimate as

$$\begin{aligned} \widehat{\text{VCE}}(x_i \mid S_x = v_t) &= \frac{1}{n} \sum_{j=1}^{n} \big(\overline{y}_{tj} - \overline{y}_t\big)^2 \\ &- \frac{1}{nr^2} \sum_{j=1}^{n} \sum_{k=1}^{r} \big(y_{tjk} - \overline{y}_{tj}\big)^2 . \end{aligned} \tag{6–12}$$

When averaged over (equally probable) sites, the PVCE is obtained as

$$\begin{aligned} \widehat{\text{PVCE}}(x_i; S_x) &= \frac{1}{sn} \sum_{t=1}^{s} \sum_{j=1}^{n} \big(\overline{y}_{tj} - \overline{y}_t\big)^2 \\ &- \frac{1}{snr^2} \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} \big(y_{tjk} - \overline{y}_{tj}\big)^2 . \end{aligned} \tag{6–13}$$

6.3.3 Partial and Incremental Partial Correlation Ratios Estimates

The estimate of the partial correlation ratio is formed as the ratio

$$\widehat{\eta}_p^2 = \widehat{\mathrm{PVCE}}(x_i; S_x) \Big/ \left\{ \frac{1}{snr^2} \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{tjk} - \overline{y}_{tj}\right)^2 \right\}. \qquad (6–14)$$

The denominator is the estimate of the residual prediction variance adjusted for x_i and S_x. The estimated incremental partial correlation ratio, based on the prediction variance, is given by

$$\widehat{\eta}_{\mathrm{inc}}^2 = \widehat{\mathrm{PVCE}}(x_i; S_x) / \widehat{V}[y] \, , \qquad (6–15)$$

where the prediction variance estimator is Equation 6–2. That estimator of prediction variance is preferred to one from the sample used to estimate the PVCE when the number of sites is small and particularly when $s = 1$.

6.3.4 Conditional Correlation Ratio Estimate

At each site t, the conditional VCE and correlation ratio can be used as local importance indicators. Local in this sense refers to the fixed value of the inputs in S_x. The conditional R_a^2 estimate is given by

$$CR_a^2 = \left\{ r \sum_{j=1}^{n} \left(\overline{y}_{tj} - \overline{y}_t\right)^2 - \frac{1}{r} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{tjk} - \overline{y}_{tj}\right)^2 \right\} \Big/ \left\{ \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{tjk} - \overline{y}_t\right)^2 \right\}, \qquad (6–16)$$

which is Equation 6–5 applied to site t. The average conditional R_a^2, weighted by the estimate of the prediction variance at site, is the estimate of the incremental partial correlation ratio.

6.4 Estimation for Arbitrary Subsets of Inputs

Methods for analysis of an arbitrary single input x readily generalize to methods for an arbitrary subset S_x. With the inputs partitioned into two disjoint subsets S_x and S_x^c, the objective of analysis is the importance of S_x. An LHS for S_x is combined with an LHS for S_x^c to allow estimation relative to the entire subset S_x and, possibly, S_x^c. If a replicated LHS is used for S_x^c, importance of its input components can be assessed as described in the Section 5.

6.4.1 Sample Design

The sample design for estimating variance components for subsets S_x of arbitrary size is essentially the one in the Section 6.2 with S_x playing the role of x and S_x^c playing the role of S_x. The number of sites s is the sample size for S_x and can be approximately the same as the number of intervals n used for individual inputs. The sample size for S_x^c is also n, and only $r = 1$ replicate is required. The sample design at site t is given by Equation 6–11, where v_t represents the values of the inputs in S_x at the site and $\boldsymbol{D}_{(S_x)}$ represents the sample on the inputs S_x^c, the same sample values of which are used at each site.

6.4.2 VCE and Correlation Ratio Estimates

The VCE for S_x is estimated as in Equation 6–3 for an individual input by

$$\widehat{\mathrm{VCE}}(S_x) = \frac{1}{s} \sum_{t=1}^{s} \left(\overline{y}_t - \overline{y}\right)^2 - \frac{1}{sn^2} \sum_{t=1}^{s} \sum_{j=1}^{n} \left(y_{tj} - \overline{y}_t\right)^2,$$

where $\overline{y}_t$ is the sample average at site t. The correlation ratio is estimated as in Equation 6–5 as

$$R_a^2(S_x) = \left\{ n \sum_{t=1}^{s} \left(\overline{y}_t - \overline{y}\right)^2 - \frac{1}{n} \sum_{t=1}^{s} \sum_{j=1}^{n} \left(y_{tj} - \overline{y}_t\right)^2 \right\} \Big/ \left\{ \sum_{t=1}^{s} \sum_{j=1}^{n} \left(y_{tj} - \overline{y}\right)^2 \right\}.$$

The symmetry of the sample design supports estimation of the VCE for S_x^c if it is statistically independent of S_x. In that case, the VCE for S_x^c is given by

$$\widehat{\mathrm{VCE}}(S_x^c) = \frac{1}{n} \sum_{j=1}^{n} \left(\overline{y}_{.j} - \overline{y}\right)^2 - \frac{1}{ns^2} \sum_{j=1}^{n} \sum_{t=1}^{s} \left(y_{tj} - \overline{y}_{.j}\right)^2,$$

where $\overline{y}_{.j}$ is the sample average at site j for the inputs in S_x^c. The reason the VCE for S_x^c cannot be estimated if S_x and S_x^c are not statistically independent is that the distribution of S_x^c given S_x may be different at different sites.

6.5 Regression Interpretations

Relationships between variance components, conditional expectations, and a general linear regression model can be seen by way on the example shown in Figure 6.1. In the figure, the prediction data from an rLHS are plotted against an individual input. All $N = n \times r$ values are plotted on the y-axis and again above the x-values to which they correspond. There are n groups of data corresponding to the n distinct values of x and n parameters in a general linear model. Within each group there are r y-values corresponding to the r replicates of the LHS. The N values on the y-axis yield the total sum of squares (SST) and correspond to the prediction variance $V[y]$. These values are partitioned into n groups with n mean values (not indicated in the figure). The group means, $\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_n$ correspond to n conditional expectations, as a function of x, and to n parameter estimates in the linear model. The means yield the between-group sum of squares (SSB) which corresponds to the VCE. From a regression perspective, the group means are the predicted values whose sum of squares is the regression sum of squares. The r values within each group correspond to regression residuals when compared with the group mean. The sums of squares about the group means form within-group sum of squares (SSW) corresponding to the residual variance component. The more important the input is, the larger is the between-group variability reflected in larger variability of group means and, at the same time, smaller residual or within-group variability. The residual or within-group variability is due to all of the other inputs. The multiple correlation coefficient $R^2 =$ SSB/SST is a measure of the goodness of fit of the regression and corresponds to the correlation ratio $\eta^2 =$ VCE$(x)/V[y]$.

6.6 Summary of Formulas for Estimation

The $N = n \times r$ observations $\{y_{jk}, j = 1, \cdots, n$ and $k = 1, \cdots, r\}$ are from an LHS of size n replicated r times. They are labeled on j to correspond to the n distinct values of an input, say, x_i.

Prediction Variance. Prediction variance is a measure of the uncertainty in y due to uncertainty in inputs x. The prediction variance is estimated in Equation 6–2 as

$$\widehat{V}[y] = \frac{1}{nr}\sum_{j=1}^{n}\sum_{k=1}^{r}\left(y_{jk} - \overline{y}\right)^2 .$$

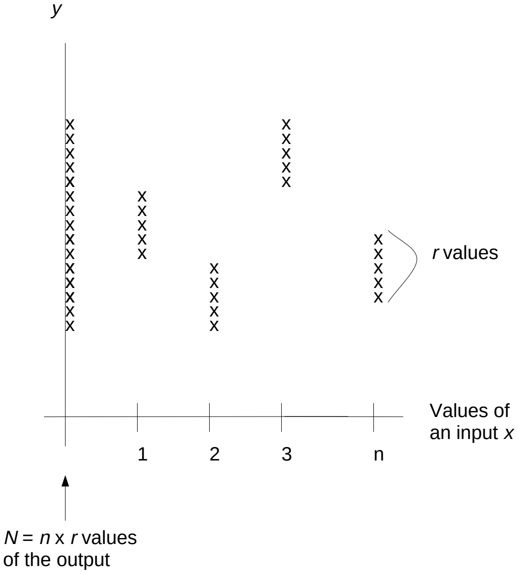


Figure 6.1 The r replicates of an LHS design for one input

VCE — VCE(x_i). An indicator of the importance of x_i is the VCE and corresponding correlation ratio. The VCE for x_i given by Equation 5–2 is

$$\text{VCE}(x_i) = V[E(y \mid x_i)] .$$

It is estimated in Equation 6–3 as

$$\widehat{\text{VCE}}(x_i) = \frac{1}{n}\sum_{j=1}^{n}\left(\overline{y}_j - \overline{y}\right)^2 - \frac{1}{nr^2}\sum_{j=1}^{n}\sum_{k=1}^{r}\left(y_{jk} - \overline{y}_j\right)^2 .$$

Correlation Ratio — η^2. The correlation ratio compares the size of the VCE with that of the prediction variance. The correlation ratio for x_i from Equation 5–5 is

$$\eta^2 = V[E(y \mid x_i)]/V[y] = \text{VCE}(x_i)/V[y] .$$

It is estimated in Equation 6–4 as

$$\widehat{\eta}^2 = R_a^2 = \widehat{\text{VCE}}(x_i)/\widehat{V}[y] .$$

In the formulas for the VCE and correlation ratio just presented, the single input x_i can be replaced by a subset of inputs S_x. In that case, the VCE and correlation ratio would be indicators of the importance of the subset of inputs. The sample values $\{y_{jk}\}$ would be labeled in j

to correspond to the distinct sample values (sites) of S_x. There may be only one such value, the vector of medians, or there may be a sample of n values from an LHS.

Important inputs are selected sequentially, in a manner similar to step-up regression. Subset S_x represents the subset of inputs selected so far and S_x^c represents the remainder of the inputs. The importance of the additional inputs, say, x^* in S_x^c is to be assessed. The sample observations $\{y_{tjk}, t = 1, \cdots, s$ and $j = 1, \cdots, n$ and $k = 1, \cdots, r\}$ are labeled on j to correspond to the n distinct values of the input x^* under consideration. The index t labels the different sites for S_x, and k, as before, indexes replicates of the LHS of size n on S_x^c.

PVCE — PVCE$(x^*; S_x)$. An indicator of the importance of the additional input x^* beyond that of the subset of inputs S_x is the PVCE and corresponding correlation ratios. The PVCE for x_i adjusted for S_x is given in Equation 5–10 as

$$\text{PVCE}(x^*; S_x) = E(V[E(y \mid \{x^*, S_x\}) \mid S_x]) \, .$$

It is estimated in Equation 6–13 by

$$\widehat{\text{PVCE}}(x_i^*; S_x) = \frac{1}{sn} \sum_{t=1}^{s} \sum_{j=1}^{n} \left(\overline{y}_{tj} - \overline{y}_t\right)^2 - \frac{1}{snr^2} \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{tjk} - \overline{y}_{tj}\right)^2 .$$

Partial Correlation Ratio — η_p^2. The partial correlation ratio compares the size of the PVCE with that of the residual prediction variance after adjustment for S_x. It is given in Equation 5–12 as

$$\begin{aligned} \eta_p^2 &= E(V[E(y \mid \{S_x, x^*\}) \mid S_x])/E(V[y \mid S_x]) \\ &= \text{PVCE}(x^*; S_x)/\text{Residual}(S_x^c; S_x) \, . \end{aligned}$$

The partial correlation ratio for x^* adjusted for the subset of inputs S_x is estimated by in Equation 6–14 by

$$\widehat{\eta}_p^2 = \widehat{\text{PVCE}}(x_i^*; S_x) \Big/ \left\{ \frac{1}{snr^2} \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} \left(y_{tjk} - \overline{y}_{tj}\right)^2 \right\} .$$

Partial Incremental Correlation Ratio — η_{inc}^2. The partial incremental correlation ratio compares the size of the PVCE with that of the (full) prediction variance. It is given in Equation 5–13 as

$$\begin{aligned} \eta_{\text{inc}}^2 &= E(V[E(y \mid \{S_x, x^*\}) \mid S_x])/V[y] \\ &= \text{PVCE}(x^*; S_x)/V[y] \, , \end{aligned}$$

and is estimated from Equation 6–15 as

$$\widehat{\eta}_{\text{inc}}^2 = \widehat{\text{PVCE}}(x_i; S_x)/\widehat{V}[y] \, .$$

The $\widehat{V}[y]$ estimate of prediction variance from the original sample can be used as a better estimate than one available from a small number s of sites.

Conditional Estimates. When there is only $s = 1$ site in estimates of the PVCE, partial correlation ratio, and partial incremental correlation ratio, the estimates are more properly called estimates of the conditional VCE, conditional correlation ratio, and conditional incremental correlation ratio, conditioned on $S_x = s_x$.

7 STEPS IN UNCERTAINTY ANALYSIS

This section provides a general overview of uncertainty analysis procedures; the steps are not tied to particular methods. The overview is not exhaustive nor does it address all possibilities likely to be encountered in a large variety of applications or, even, in the sample applications in this report. The steps are divided into four parts: problem definition, screening, validation, and methods sensitivity and diagnostic tests.

7.1 Preliminary Considerations and Problem Definition

Objectives of the analysis of the prediction y from the model $m(\cdot)$ can be stated succinctly as (1) to quantify prediction uncertainty and (2) to quantify the importance of inputs with regards to prediction uncertainty. The statement may be misleading in its simplicity because it does not mention all of the specifications and restrictions which apply. For example, early in the course of performing the analysis it becomes clear that prediction uncertainty for $y = m(x)$ cannot refer to *every* prediction using $m(\cdot)$, but only to the specific application under study, as quantified by the inputs x and their probability distribution f_x. Therefore, the appropriateness of the representation of reality reflected by f_x must be duly assessed. The assessment extends both to the range V—including range dependencies—of the inputs and to the form of the probability distribution. It is important to remember that uncertainty analysis is relative to the uncertainty triple $(f_x, V, m(\cdot))$.

7.1.1 Definitions of Model Predictions and Selection of Model Outputs

The actual model (computer code) outputs to be recorded for each model run are specified. The model predictions of interest might be outputs or they might be derived from outputs. For example, an output might be deposition on a spatial grid at several time steps. The corresponding prediction of interest might be integrated deposition at each time step or just total deposition. For simplicity of presentation, the rest of this section considers a single prediction y.

7.1.2 Identification and Specification of Model Inputs

Inputs describe the application both in terms of initial conditions of the scenario and in terms of the process dynamics modeled by $m(\cdot)$. All relevant inputs, whether they are called input variables, parameters, or data, and whether they come from external input files or are hard-wired in the code, are identified as being part of the context of the uncertainty study. Some of the variables are likely to be set to fixed values and not changed at all. These variables are thought of as being assimilated into the function $m(\cdot)$. The remaining variables, those whose values are assumed to have input uncertainties, are the ones denoted by x and called inputs.

7.1.3 Assignment of Probability Distributions

For each input, limits on the range of values are specified. Although narrow limits might be appropriate for a preliminary uncertainty analysis, they should be wide enough to provide adequate coverage in the anticipated input space. On some occasions, however, accurate bounds may be necessary, particularly when $m(\cdot)$ is sensitive to values at a boundary.

The joint range of values of some sets of inputs may exhibit dependencies and not be the product of their individual ranges. A common example of this behavior is where the value of one input should not exceed that of another. All such joint range dependencies are specified.

Probability distribution f_x is constructed in parts: individual or marginal distributions for inputs that are statistically independent, and joint distributions for those subsets of inputs that are dependent. Each subset of dependent inputs falls into one of two kinds: the joint range of values is the product of the individual ranges or it is not. After all specification of ranges of values, the forms of the probability distributions are determined. For preliminary uncertainty studies, simple distributions for f_x, like the uniform and beta for finite ranges and the normal and exponential for infinite ranges, and their logarithmic cousins are often used for convenience' sake. In any event, an examination of sensitivity of conclusions to the choice of f_x can be both informative and necessary as part of scientific investigation.

7.1.4 Construction of the Base Case Set of Runs

The specification of model outputs, inputs, ranges of values, input distributions, fixed input variables, and their values provides the information needed to generate a sample of model computer runs from which the distribution of y can be estimated. For independent inputs, an LHS of appropriate size (n) is used. For inputs that are not independent, the joint distribution is sampled in any appropriate way, including the possibility of a stratified sample.

The question of sample size is always present and almost never answered satisfactorily. Several points bear consideration. First, interest lies in estimation of the density function f_y, and to that end, there are many possible estimators. Statistical literature might provide an evaluation of properties of the sampling plan used (LHS) with respect to mean square error, say, of the density estimator. Such an evaluation is beyond the scope of this report. One approach to an empirical evaluation is to try several sample sizes, say, 100, 500, and 1000, and several samples of each size. An examination of different estimates obtained from the samples and with different settings of parameters in the density estimation algorithm will guide one to a reasonable choice. It is supposed that a very large number of computer runs can be made, and so computer resources do not pose a conceptual limit on the analysis. When computing resources are limited, either in time or money, compromises will, undoubtedly, occur.

It is supposed that the base case set of runs consists of an LHS of size 100, and from those data an acceptable estimate of the density function of the prediction y is constructed. The analysis continues with the identification of those inputs and input subsets that are important with respect to prediction variance.

7.2 Sequential Screening

The analysis proceeds from the base case to construction of candidate subsets of important inputs. The measure of prediction uncertainty used is prediction variance and the measure of importance could be the correlation ratio. The process is called screening because the subsets are only candidates to be tested or “validated” before being accepted as important. Screening is intended to allow for a series of trial selections of important input subsets which are later validated through a few comprehensive tests. Several alternatives methods to indicate importance are available. The sampling plans LHS and rLHS support both variance estimation for the correlation ratio and also regression methods which include partial correlation. Both of these methods can be used simultaneously to construct candidate subsets.

Most useful would be the determination of candidate subsets of size 1, 2, 3, and so forth. The product of this phase of analysis is lists of input subsets $C^s = \{C^s_j,\ j = 1, 2, \cdots\}$ representing candidate subsets of size s inputs each. Normally, interest might reside in best subsets of size 1, 2, 3, and so forth up to some (small) number of inputs that can be said to account for essentially all of the uncertainty in y. In mathematical terms, the process finds for each size s those subsets j^* for which $V[y] -$ VCE$\big(C^s_{j^*}\big)$ is relatively small, subject to sampling variability. Because of time for computation, some analyses will proceed as directly as possible to the smallest acceptable subset of important inputs. The full sequential technique is used in the application in Section 8 and the abbreviated one is used in the application in Section 9.

When y is really several predictions, one might perform several screening exercises in parallel by constructing candidate subset C^s to be the superset of candidates, those inputs considered important for at least one prediction. This process has a drawback if there are multiples stages in screening, as described in Section 9. Namely, the result is really the identification of the inputs which are unimportant for any of the outputs. If identification of important inputs for each output is necessary, separate analyses for each output may be necessary.

7.3 Validation and Diagnostic Testing

Validation and diagnostic testing provide independent evaluation pointing towards confirmation of the importance of inputs selected with the screening procedures. Let S_x denote the set of inputs to be validated and S^c_x the remaining inputs. Validation consists of two complementary steps. The first step in validation is to examine the conditional prediction distributions when the supposed important inputs S_x are held fixed. If these distributions, independent of the conditioning value s_x, reflect a substantial reduction in uncertainty as compared with the unconditional prediction distribution, then S_x is confirmed as important relative to prediction uncertainty. The sample of conditioning values $\{s_{x1}, s_{x2}, s_{x3}, \cdots\}$ at which S_x is fixed must be adequate to cover the range of S_x. It is not possible to state a sample size generally adequate for all analyses.

The second step in validation examines the conditional distribution of y when the supposed unimportant inputs S_x^c are held fixed. The conditional distribution of y is expected to look very much like the unconditional prediction distribution independent of the value of S_x^c. These two validation steps follow from the characterizations of importance and prediction uncertainty in Section 4.

Finally, diagnostic testing is meant to describe the examination of the data generated during screening and validation. Simple procedures such displaying output values y, or scatter plotting y versus x, or displaying sample standard deviations for candidate subsets or for different sites for fixed subsets can all point out important relationships and behaviors which go undetected by summary statistics used in screening.

7.4 Summary of Steps in Uncertainty Analysis

Uncertainty analysis consists of two parts: preliminary analysis and final analysis. The main difference between the two lies in the probability distributions of the inputs. In a preliminary analysis, approximate distributions like the uniform and loguniform that are easy to work with are matched (fit) to the range, mean, percentiles, and other information about the inputs. Results of the analysis may be tested for sensitivity to changes in distributions. In a final analysis, best estimates for critical input distributions are used. A typical sequence of steps appears below.

(1) Identify and describe all potential parameters or input variables. They fall into three categories: those relating to the numerical algorithms, those describing phenomenology or mechanics of the process being modeled, and those describing the event or scenario being studied.

(2) Identify any inputs which will not be further considered and state how they will be assigned values.

(3) Identify any subsets of inputs that cannot be varied independently.

(4) Choose ranges of variation for those inputs that can be varied independently.

(5) Define domains for each subset of inputs that cannot be varied independently.

(6) Assign uniform or loguniform distributions to independent inputs.

(7) Define appropriate joint distribution functions for dependent subsets of inputs.

(8) Obtain base case sample where all inputs vary.

(9) Determine important subsets of inputs:

   (a) Initial stage analysis. Important inputs are determined for each output. Those inputs not in any of the lists are deemed unimportant.

   (b) Subsequent analyses. Separate sequential screening is done for each output to determine important inputs. As before, those inputs not in any of the lists are deemed unimportant.

   Sequential analyses serves several purposes and can produce a more complete subset of important inputs the more the model deviates from linearity.

(10) Perform suitable validation and diagnostic testing:

   (a) Unless a single subset of important inputs for all outputs is to be identified, each model output should be analyzed independently relative to its own subset of important inputs.

   (b) If important subsets do not sufficiently account for prediction uncertainty, continue with the sequential input selection in (9).

   (c) Data from the analysis is examined to reveal any previously undetected relationships and behaviors.

(11) Determine final probability distributions for important inputs, and assign the same preliminary distributions to the unimportant ones.

(12) Choose among alternative submodels.

(13) Repeat (8)–(10).

(14) Examine sensitivity of results to perturbation of input distributions. This step is another uncertainty analysis in itself, where the “inputs” define the real input distributions. Whether a formal or informal analysis is carried out is a matter of choice.

(15) Continue if any corrective actions appear necessary.

8 ANALYSIS APPLICATION I

In the discussion to follow, the parts of an analysis presented in the previous section are applied to a model which predicts flow of material in an ecosystem. The details of application are intended to serve as a guide to a simple uncertainty study. The analysis follows McKay and Beckman (1994b) where important inputs are screened—tested for importance—in stages: top single inputs, top pairs, top triples, and so forth. The top 10 or so candidates at each stage move on to the next stage to be augmented by an additional input. This type of analysis requires an extensive number of runs, so it may not be appropriate for all models. The second analysis application (Section 9) uses a modification of the procedure for longer-running models.

8.1 Problem Definition

The model $m(\cdot)$ is a compartmental model of an ecosystem. The flow of material among the several compartments, indicated in Figure 8.1, is described by a set of linear differential equations which relate concentrations in compartments as functions of time. Although eight output compartment concentrations are outputs calculated by the model, only the concentration in compartment C3 at a large value of time when the system is in equilibrium is considered in this analysis. The inputs x are 84 coefficients comprising initial conditions and transfer coefficients. The uncertainty analysis is required because of uncertainty in appropriate values of the 84 inputs.

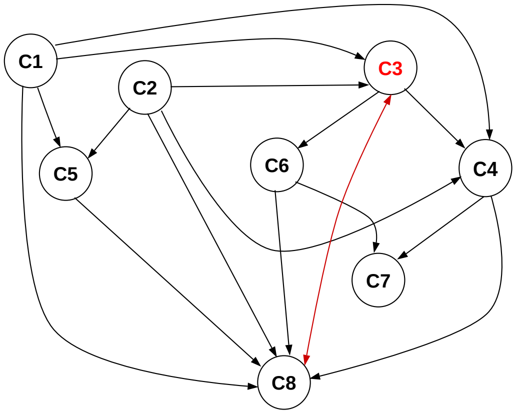


Figure 8.1 Compartmental model

The purpose of the analysis is twofold: first, to obtain a preliminary estimate of the variability of prediction due to input variability and, second, to supply guidance for refining uncertainty limits for input values. Literature review and expert judgement provide absolute ranges and best estimate values for each input. Because the analysis is preliminary, only minimal effort is expended to quantify shapes of probability distributions on the ranges or in investigating and representing statistical dependencies among inputs. Independent uniform probability distributions are used for all inputs except those that vary over several orders of magnitude, for which loguniform distributions are used. In summary, the study considers

- y, the concentration in compartment C3 at equilibrium
- x, a vector of 84 inputs which are parameters in differential equations that govern the concentrations
- f_x, a joint, independent uniform or loguniform probability distribution for the inputs

An assumption motivating the analysis is that reducing the uncertainty in a subset of the inputs reduces the uncertainty in y. Whether or not the assumption is true in this case will be investigated by examining conditional distributions of y when important inputs are held fixed. The assumptions of independent uniform distributions of the inputs is not examined in the application, although that would be necessary in a complete uncertainty analysis. Finally, the model runs very quickly and so there are essentially no limitations on the number of computer runs that can be made.

8.1.1 Base Case Sample

A base case sample of size 250 is constructed as described in Section 6.1.1 for an rLHS of size $n = 25$ with $r = 10$ replicates. Construction of the sample design begins with the 25 row × 84 column matrix $\boldsymbol{D}_0$ corresponding to an LHS of size 25. The replicates are formed from that matrix by randomly permuting its columns 10 times to form to the full 250 × 84 design matrix

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_{10} \end{bmatrix} .$$

The construction points out that the same 25 values for each input appear in each of the 10 replicate design matrices and that the replicate design matrices differ in the input combinations designated by the rows of the matrices.

8.1.2 Prediction Distribution

The estimate of prediction density function f_y obtained from the base case sample of size N = 25 × 10 = 250 runs is shown in Figure 8.2. The density function was estimated using the function "density" in the S Language (Becker, Chambers, and Wilks, 1988) from the S–PLUS software (Statistical Science, 1991). Although the mode of the distribution is about 10 and there is a lower bound of 0 on concentration, values larger than 100 are very likely. The long tail of the distribution extends beyond 400 and shows that y has a wide range of variation. Because of the nonsymmetric shape of f_y, simple measures like the mean value (168), the median (26), and the standard deviation (400) are inadequate as full descriptions of the probability distribution of y. In fact, the range of the data used in estimation is 0.02 to 7700. Therefore, the effect on uncertainty of reducing prediction variance is observed better and more completely in the (estimated) density function itself.

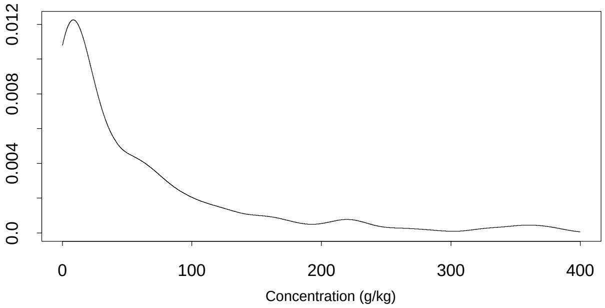


Figure 8.2 Estimate of prediction density f_y

In the base case data, there are 10 very large values of y (on the order of 7000 g/kg) coming from the tail of its distribution. These outliers have undue influence on the sample variance used in importance measures. The situation is alleviated by use of the rank transform of y when doing input screening. For the base case data, the 250 values of y are ordered from smallest to largest. The smallest value is replaced by 1, the next smallest by 2, and so forth to the largest value which is replace by 250. There is nothing particularly optimal about the rank transformation in this application: the logarithmic transformation would be another acceptable choice. Although the rank-transformed data are used in screening when selecting potentially important inputs, the density functions of y examined in validation are calculated with the original data values.

8.2 Sequential Screening Procedure

The correlation ratio (Eq. 5–5) and other variance ratios are used as importance indicators for screening. Nevertheless, the concurrent use of the partial rank correlation coefficient is a reasonable practice. Subsets of inputs (S_x) are evaluated in stages. Those whose importance indicators are large enough become the candidate or top subsets in the stage. In Stage 1, base case data are examined for individual inputs. The top 10 or so inputs are designated the Stage 1 candidates and become the first elements of subsets of size 2. The top 10 or so of these become the first 2 elements for subsets of size 3, and so forth. At each stage h, subsets of size h are examined, and those most promising—the 10 or so—become candidate subsets in the list C^h. There is no reason that candidate lists be constructed by increments of one input. For example, it is reasonable to proceed from the list of candidate individual inputs C^1, to examination of subsets of size 3, say. This strategy might be taken when three inputs stand out as dominating prediction variance, strongly suggesting that any important subset would contain all three of them. However, when inputs are dependent it may not be necessary that they all be selected as part of the subset of important inputs.

The transition from one stage to the next is explained for Stage 1 to Stage 2. The transition from arbitrary stages h to $h' > h+1$ uses a different procedure and is explained in a subsequent section. In general, the procedure requires a large number of computer runs, so it may not be practical for all models. Reasonable modifications include selection of only the top candidate at each stage and its subsequent augmentation by several inputs instead of just one at a time—the approach used for the second application. Discussion of details of the analysis continues.

8.2.1 Stage 1

For each of the inputs, the ranked values of y are ordered and relabeled to correspond to the ordered values of that input and are used in the computations of the VCE and R_a^2 (or R^2) from Equations 6–3 and 6–5. The denominator of R_a^2 is the same for all inputs because it is independent of the ordering of y-values.

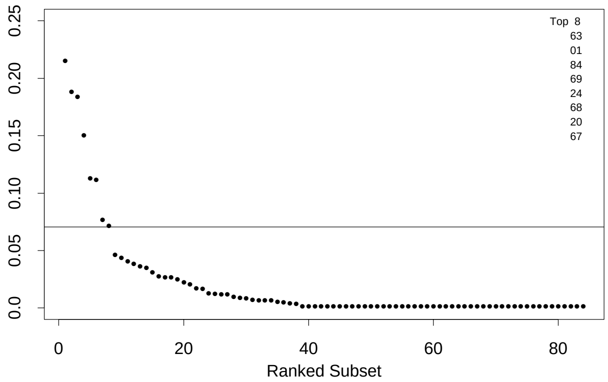


Figure 8.3 Ordered R_a^2 for single inputs

The 84 values of R_a^2 for each input are plotted in Figure 8.3. The values are plotted from largest to smallest, and negative values are plotted as zeros. The largest 8 values correspond to inputs in the list of candidates $C^1 = \{63, 1, 84, 69, 24, 68, 20, 67\}$. The largest few of these correspond to only about 20% of variability accounted for, indicating the contribution of any individual input to prediction variance is less than 20%. Moreover, because of the sample-to-sample variability expected in the values of R_a^2, it is difficult to point to any of the larger values as being significantly different from others.

The R_a^2-values for the inputs in C^1 appear large enough to be set off from the rest. However, since no individual input subset S^1 of size 1 dominates uncertainty for y, the analysis continues with identification of candidate input subsets S^2 of size 2.

Before proceeding to Stage 2, the effect of input 1 is investigated in more detail to provide insight into the analysis process. The effect on uncertainty in y of x_1, whose R_a^2 is about 0.20, is shown through two sets of conditional distributions. This part of the analysis has the flavor of the validation step: the effect on the uncertainty in y of fixing input 1 at different values is examined. When input 1 is fixed, the variability in y is due to the other 83 inputs varying. Ten values of x_1 are selected using LHS to provide a sample of x_1 which spans its range. For the other 83 inputs, an LHS of size 250 is constructed which samples their 83–dimensional space. At each point in the x_1 sample, 250 model runs are made using that value and the LHS-250 for values of the other inputs. The 250 output values y are used to estimate the conditional density of y given x_1 fixed. These 10 conditional densities are plotted in Figure 8.4. The densities show the effect of changing the fixed value of x_1 on the distribution of y, and offer the interpretation to the R_a^2 value of about 0.20. Namely, while x_1 alters the distribution of y, it alone does not seem to be able to significantly reduce the prediction variance.

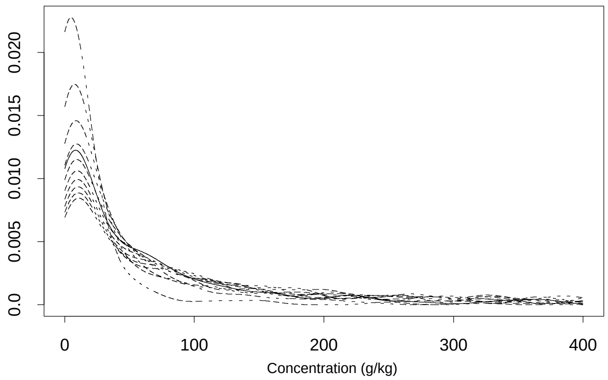


Figure 8.4 Conditional densities of y for 10 values of $\{x_1\}$

To complete the examination related to x_1, the effect of the complement subset of 83 inputs is examined in the same way. An LHS of size 10 is used to select 10 values of the 83-tuple of other inputs; an LHS of size 250 is used to select 250 values of x_1. For each value of the 83-vector, the conditional density of y is estimated from the runs with the 250 values of x_1. The densities are plotted in Figure 8.5. The patterns in the figure show two things. First of all, each curve indicates the extent of the variability in y caused by x_1, because only x_1 varies for each density. Second, as expected, the curves indicate existence of important inputs among the 83 by the differences among the 10 densities. The next stage in the analysis looks for important inputs from the 83.

8.2.2 Stage 2

Stage 2 denotes the construction of subsets of size 2. It generalizes to transitions from subset size h to $h + 1$ and is discussed accordingly. The possible subsets considered are all of those that include a member from the list C^1 identified in Stage 1. That means, for example, x_1 is allowed to pair with any of the other 83 inputs, but x_2 can only pair with the inputs in C^1. It is true that there may be important pairs of inputs not containing any of the inputs in C^1 and which may not be identified in this stage. There are 3486 possible pairs of which only 8 × 83 = 664 are to be examined because only 8 inputs were selected as candidates in Stage 1. Although special algorithms

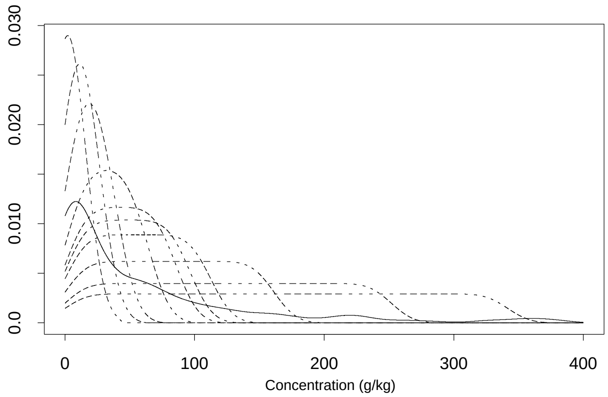


Figure 8.5 Conditional densities of y for 10 values of $\{x_2, \cdots, x_{84}\}$

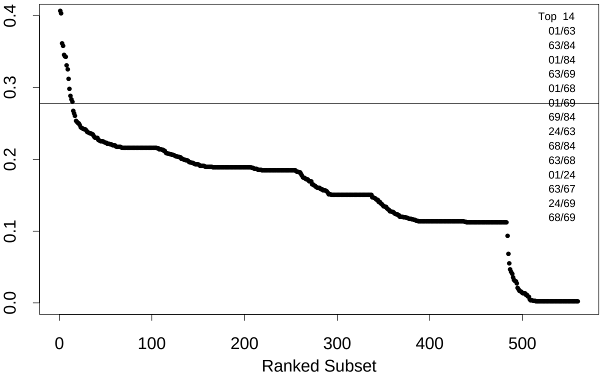


Figure 8.6 Ordered R_a^2 for pairs of inputs

exist to select optimal subsets in variable selection in regression, step-up, and step-wise procedures are still used. The procedures described here are similar to the heuristic procedure of step-up variable selection. Heuristic procedures and investigator intuition must suffice until approximate bounds on optimal VCE are developed.

Stage 2 calculations are explained for x_1, on one of the candidates from Stage 1, before presentation of complete results for Stage 2. With x_1 set to a fixed value, an analysis like the one for Stage 1 for the remaining 83 inputs can be performed. Thus, the other 83 inputs can be screened for important inputs conditioned on the value of x_1 using the conditional VCE and R_a^2. If the calculations are carried out at several "sites" for x_1 and suitably averaged, the expected value of the VCE—the partial VCE adjusted for x_1—is estimated for use in the computation of the partial correlation ratio adjusted for x_1. Thus, Stage 2 is essentially just Stage 1 at a sample of sites for x_1. Finally, the VCE and correlation ratio for each full subset of size 2 made with x_1 can be calculated from Equation 6–10.

The design for Stage 2 has two components: a design matrix for the values of x_1 and another for the values of the other inputs. Using Taguchi terminology (Taguchi, 1986), the design on x_1 would be called the outer array and the one on the other inputs the inner array. The full design is the product of the two. To provide some continuity of sample y-values for comparison purposes—so that changes observed are less likely to be due to sample-to-sample variations seen in independent samples—modifications of the original base case design matrix **D** are used for designs in all stages. At each of s sites, 250 runs are made using the base case design matrix **D** with the column corresponding to x_1 replaced by a column of the constant site value of x_1. In this application, $s = 4$ site values are used for all inputs in C^1. The values are the 12.5, 37.5, 62.5, and 87.5 percentiles of the distributions of the inputs.

For each input, the conditional VCE and R_a^2 are estimated at each site from Equations 6–12 and 6–16. The estimates are combined to form the PVCE and R_a^2 adjusted for x_1 as in Equations 6–13 and 6–14. Finally, the R_a^2 for each 2-input subset is formed as the sum of the VCE(x_1) estimated in Stage 1 and the PVCE estimate in Equation 6–13. Similar calculations for the rest of the inputs in C^1 complete the computations.

Ordered values of R_a^2 for 2-input subsets are presented in Figure 8.6. A natural grouping like the one in Stage 1 is not as apparent, so a somewhat arbitrary cutoff at 14 pairs is selected in the figure. Of interest, however, is that all of the top 14 pairs are composed of inputs from C^1 and that maximum R_a^2-values are about 0.40. This suggests that minimal subsets will include several inputs. The wavy pattern in Figure 8.6 comes from strings of subsets having common members.

The candidate pairs selected in Stage 2 are indicated in Figure 8.7. The candidate list C^2 includes all 10 possible pairs from the input subset $\{1, 63, 68, 69, 84\}$ plus the 2 pairs of 63 with $\{24, 67\}$ plus the 2 pairs of 24 with $\{1, 69\}$. Incremental R_a^2-values are informative because they provide the incremental contribution of additional single inputs adjusted for the presence of the already selected input(s). Comparison of Figures 8.3 and 8.6 indicate by subtraction incremental R_a^2-values. The 5-input subset of $\{1, 63, 68, 69, 84\}$ is strongly suggested as part of a minimal important subset, so moving directly to Stage 6 with the single 5-tuple candidate $C^5 = \{(1, 63, 68, 69, 84)\}$ is indicated.

	1	63	68	69	84	24	67
1		x	x	x	x	x	–
63	x		x	x	x	x	x
68	x	x		x	x	–	–
69	x	x	x		x	x	–
84	x	x	x	x		–	–
24	x	x	–	x	–		–
67	–	x	–	–	–	–	

Figure 8.7 Candidate input pairs in C^2

The analysis for x_1 alone used (4 values of x_1) × (rLHS-25 × 10) = 4 × 250 = 1000 computer runs. To perform the same calculations for the 8 inputs in C^1 requires 8000 runs. Because the model was fast running, no consideration was given to limiting sample sizes. In other applications the number of runs might have to be reduced. By no means is it intended that 8000 is a required minimum number; it was used for convenience. In this application there may be a question of the adequacy of using only 4 points when augmenting. The important issue is whether inputs are overlooked in the screening process. The issue is addressed in the validation step.

8.2.3 Subsequent Stages

Subsequent stages evolve very much like Stage 2 from Stage 1. The candidate list C^h from Stage h consists of subsets S^h of h inputs. Each subset is sampled according to an LHS of size 4 to generate the 4 sites $\{v_1, v_2, v_3, v_4\}$. The design matrix at site t is

$$v_t \otimes D_{(S^2)} \ ,$$

meaning that the base case design is modified by replacing the original values for the subset S^h with the site values v_t. Other computations proceed as in Stage 2, except that the previously selected candidate is a multiple input subset rather than just a single input, although it is treated logically as an input variable.

Ordered values of R_a^2 for Stage 3 are presented in Figure 8.8. A somewhat arbitrary cutoff at 16 triples is selected in the figure. Of interest is that all of the triples except for one come from the 6-subset $S^6 = \{1, 24, 63, 68, 69, 84\}$, and that all six inputs in the top triples appear in C^1.

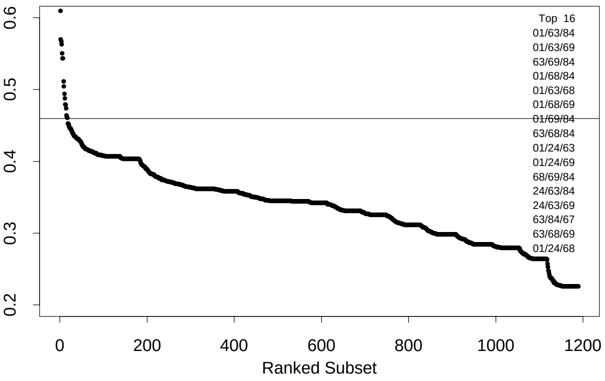


Figure 8.8 Ordered R_a^2 for triples of inputs

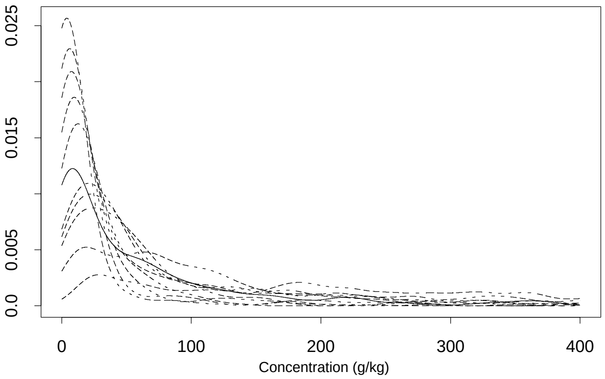


Figure 8.9 Conditional densities of y for 10 values of $\{x_1, x_{68}, x_{69}\}$

After the 16th largest R_a^2, inputs other than those in $S^6 = \{1, 24, 63, 68, 69, 84\}$ appear, and so the top 20 are not all the possible triples from S^6.

One of the top input subsets is $\{1, 68, 69\}$. The conditional prediction densities in Figure 8.9 show that the subset significantly reduces the variability in y as indicated in the tails of the densities when compared with the marginal distribution in Figure 8.2. The complementary conditional prediction densities in Figure 8.10 look very similar to those in Figure 8.9, suggesting that the analysis may be about halfway to completion.

The sequential screening procedure terminated with the selection of a single subset of 11 of the 84 inputs, which causes the long, heavy tail of the prediction distribution.

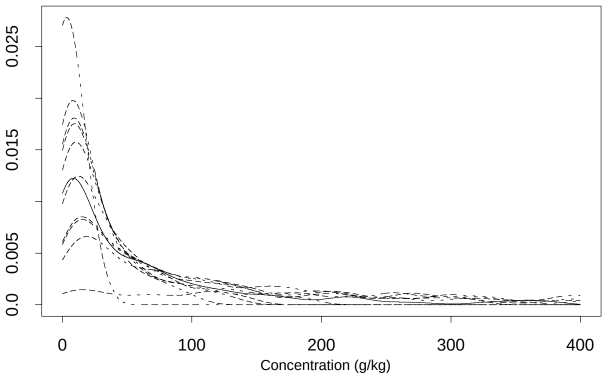


Figure 8.10 Conditional densities of y for 10 values of $\{x_2, \cdots, x_{67}, x_{70}, \cdots, x_{84}\}$

8.3 Validation

The purpose of validation is to provide confirmation that subsets identified as important do indeed control or explain prediction uncertainty. In the application, prediction variance was the criterion for screening, and so reduction in prediction variance by important subsets, as measured by the VCE, is guaranteed. Nevertheless, confirmation through examination of conditional prediction densities is required. Had the screening criterion been partial correlation, for example, reduction in prediction variance would not have been as obvious. Plots of conditional densities for smaller input subsets have already appeared. Examination of the effect of the final, 11-input subset follows.

Ten conditional prediction densities corresponding to 10 values for the subset $S^{11} = \{1, 24, 35, 48, 54, 63, 67, 68, 69, 83, 84\}$ are given in Figure 8.11. For each of the 10 densities, the effective range in y is about 100. The marginal prediction density appears in the figure for reference. Relative to it, controlling S^{11} essentially controls y.

The residual variability which causes the spread in each of the 10 densities is due to the remaining 73 inputs. Figure 8.12 shows that for each of 10 sample values of the 73 inputs, the variability in y due to S^{11} looks essentially like the prediction distribution. Thus, the objective of identifying a (small) subset of the inputs that essentially accounts for the uncertainty in y is satisfied in S^{11}. The residual uncertainty of 100 or so is not significant, in this application, relative to the full, free range of 7000. However, the analysis could continue with the identification of additional inputs which would reduce the range in y even more.

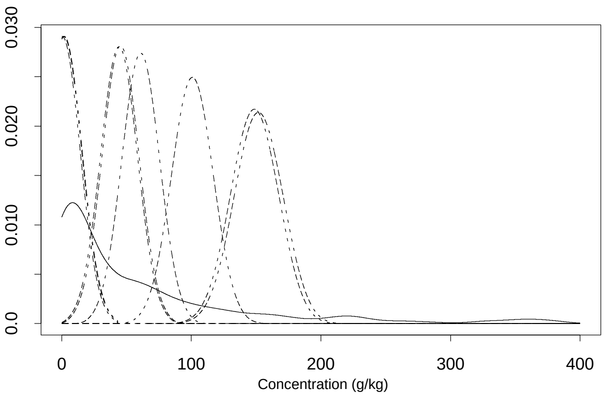


Figure 8.11 Conditional densities of y for 10 values of the important inputs

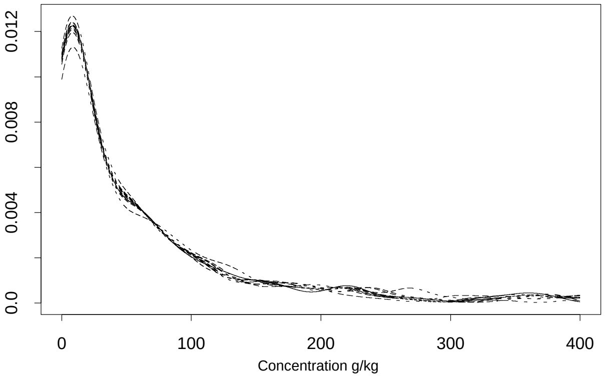


Figure 8.12 Conditional densities of y for 10 values of the less important inputs

As a final diagnostic aid, values of the square root of the residual variance (the residual standard deviation) from the candidate subsets selected in each of the 11 stages are presented in Figures 8.13 and 8.14. The plots point out two things. First of all, prediction variance is not substantially reduced beyond that achieved by the best subsets of size 4 and 5. Second, the standard deviation from rank data decreases approximately linearly with size of the best subset until it reaches its theoretical minimum indicated by the horizontal line.

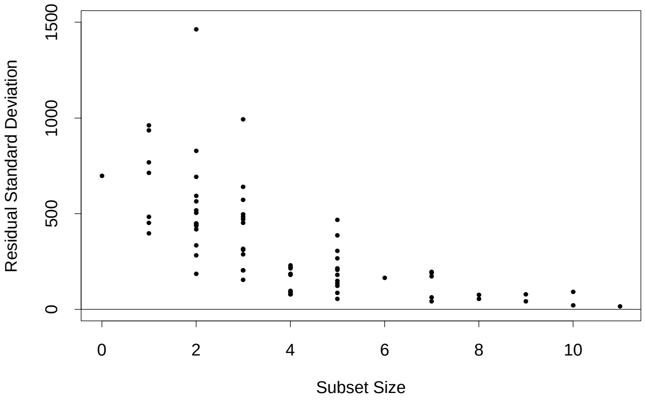


Figure 8.13 Residual standard deviations in concentration units for candidate subsets

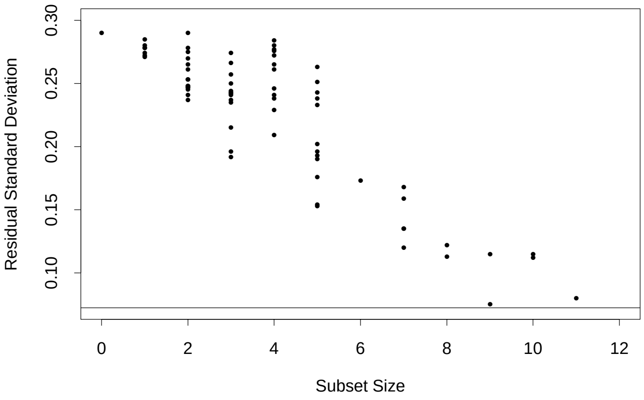


Figure 8.14 Residual standard deviations for rank-y values for candidate subsets

9 ANALYSIS APPLICATION II

The model used in this application is MELCOR Accident Consequence Code System (MACCS), described in Helton et al. (1992). The purpose of MACCS is to simulate the impact of severe accidents at nuclear power plants on the surrounding environment. In any particular application of MACCS there are likely to be many possible inputs and outputs of interest. For this application, attention focuses on 3 outputs and 36 inputs. The objective is to determine a subset of the 36 model inputs that is dominant, or important, in the sense that they are the principal contributors to prediction uncertainty. The analysis follows McKay and Beckman (1994a).

This application differs from the first (Section 8) in three ways. First of all, MACCS takes much longer to run, therefore a sequential analysis based on selection of inputs one at a time is replaced by the modified approach where several inputs are selected at each stage. Secondly, there are three model outputs rather than just one. Finally, the outputs are vector valued rather than simple scalars. The prediction $y(t)$ is given for discrete values of $t \in \{t_1 < t_2 < \cdots < t_m\}$. Therefore, the notation y means the vector of output values

$$y = (y(t_1), y(t_2), \cdots, y(t_m))^t \,.$$

9.1 Problem Definition

MACCS calculates consequences of a reactor accident at a nuclear power station whose characteristics and those of the surrounding environment are defined by inputs. Because the purpose of this section is to demonstrate methods, the inputs are identified only by number. The names of the MACCS input variables are given in the MACCS User's Guide (Chanin et al., 1990) and listed in Appendix B. The 36 inputs selected for study are only some of the 67 used for MACCS input. Therefore, their input numbers lie in the range from 1 and 67. The outputs selected for examination are Early Fatalities (the number of fatalities within 1 year of the accident), Total Cancer Fatalities, and Population Dose. MACCS is composed of submodels for source term, plume rise, atmospheric transport, dry deposition, wet deposition, evacuation, food chain transport, and dosimetry and health effects. Analysts determined plausible ranges of uncertainty for the inputs from the literature, experimental results, and submodel considerations. Because of the preliminary nature of this particular analysis, uniform and loguniform probability distributions defined on input ranges are used. Joint probability distributions or sample correlations are used for subsets of inputs that can not be treated reasonably as statistically independent. For many more details on this part of the analysis process see Helton et al. (1992).

A necessary input to MACCS is weather condition. Because weather is a random phenomenon, MACCS can be thought of as a stochastic model when weather is a sampled input. To account for the stochastic variability due to weather, MACCS computes as outputs three complementary cumulative distribution functions (CCDFs) corresponding to Early Fatalities (EF), Total Cancer Fatalities (CF), and Population Dose (PD). The CCDFs are induced by treating weather conditions at the time of the accident as a random phenomenon. Tables of weather parameters (1 year of hourly readings of wind speed, wind direction, atmospheric stability, and precipitation) are sampled repeatedly during the MACCS run to produce, in effect, a Monte Carlo estimate of the CCDF, denoted by $y(t)$. Therefore, the model "prediction" corresponding to Early Fatalities is

$$\begin{aligned} y(t) &= \text{EF} \\ &= \Pr\{\text{Number of Early Fatalities} > t\} \\ &\qquad \text{for } t = t_1 < t_2 < \cdots < t_m \,. \end{aligned}$$

Strictly speaking, for each set $\{t_1, t_2, \cdots, t_m\}$, the set $\{y(t_1), y(t_2), \cdots, y(t_m)\}$ has a joint distribution. However, it is sufficient for the analysis to examine the distributions of $y(t)$ for each t separately. For the sake of discussion, the actual values of t have been replaced by the integers 1, 2, 3, and so forth in what follows.

9.1.1 Base Case Sample

The base case sample is an rLHS with $r = 10$ replicates of an LHS of size $n = 50$. The $n = 50$ size was used because of a code requirement for generating correlated samples, as described by Iman and Conover (1982). The $r = 10$ replicates is a somewhat arbitrary number that could have been estimated by preliminary analyses. It took about 12 hours to make 500 MACCS runs. Results for EF are presented first.

9.1.2 Prediction Distribution

Each of the three outputs, EF, CF, and PD, actually has 81 prediction densities corresponding to the 81 values of t. Rather than calculating formal density estimates, which would be somewhat difficult to interpret even as a 3-dimensional plot, the prediction variability of the outputs is presented informally in a plot of the actual output calculations from the first replicate of the base case sample. The data constitute a full LHS, and so give a representative sample of model predictions. The representative data for EF in Figure 9.1 are 50 CCDFs for t from 1 to 50. The traces indicate regions of higher and lower concentration of CCDFs.

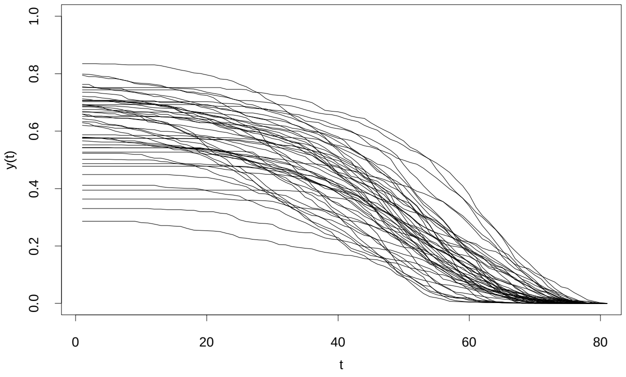


Figure 9.1 Representative $y(t)$ for EF

The representative data for CF in Figure 9.2 are 50 CCDFs for t from 40 to 81. For values of t less than 40, CF and PD are both constant at 1. The traces indicate two bands of CCDFs. The lower band contains about 20% of the data.

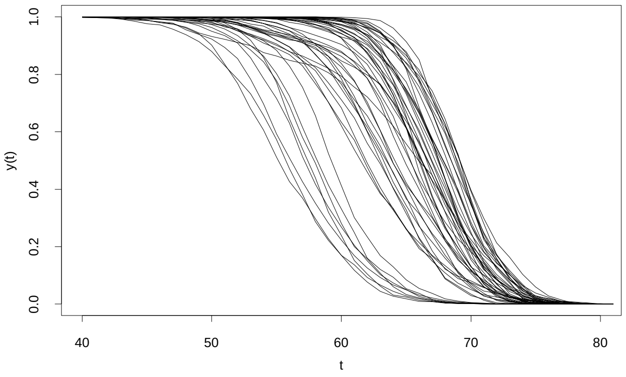


Figure 9.2 Representative $y(t)$ for CF

The representative data for PD in Figure 9.3 are 50 CCDFs for t from 40 to 50. The traces show a relatively uniform concentration of CCDFs except for one high and one low CCDF.

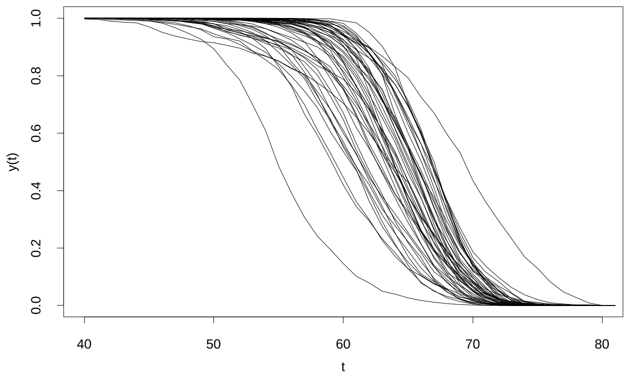


Figure 9.3 Representative $y(t)$ for PD

9.2 Sequential Screening Procedure

The selection of important inputs in this application differs from that in the first application for two reasons. First of all, there are three model outputs rather than one. Ultimately, each output is analyzed separately, although alternative approaches might have been taken. Second, each output is a vector of values rather than a simple scalar. Thus, calculations of statistics and statements about importance of inputs with respect to an output actually refer to the several "outputs" in the vector of output values.

9.2.1 Stage 1 for All Outputs

The first step in identification of important inputs is the calculation of $R^2(t)$ for each input. The $R^2(t)$ are computed with the base case sample data and are plotted in Figures 9.4–9.6. Rank-transformed y-values are used.

The standard deviation of the full base case sample, the first 50 values of which appear in Figure 9.1, is plotted as

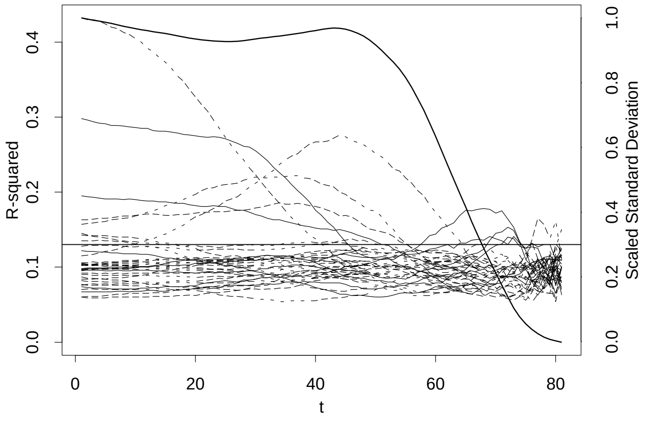


Figure 9.4 The $R^2(t)$ for 36 inputs for EF

the heavy curve ranging between 0 and 1 and indicated on the right-hand axis. The standard deviation has been normalized to a maximum value of 1, which occurs in Figure 9.4 at $t = 1$. The importance of inputs as indicated by their $R^2(t)$ is to be viewed relative to the size of the standard deviation. When the standard deviation is small, as it is for $t > 70$ or so, importance of inputs is not particularly relevant. This point is seen to be more meaningful for the smaller values of t in Figure 9.5. Finally, the horizontal line extending from 40 to 81 corresponds to the 95% critical value for R^2 and is used as a reference point for preliminary selection of important inputs. At the first stage, 10 inputs are identified as important: numbers 27, 30, 31, 33, 38, 40, 42, 50, 59, and 65. It is apparent that importance of inputs depends of the value of t. The choices made represent inputs that appear important for some values of t.

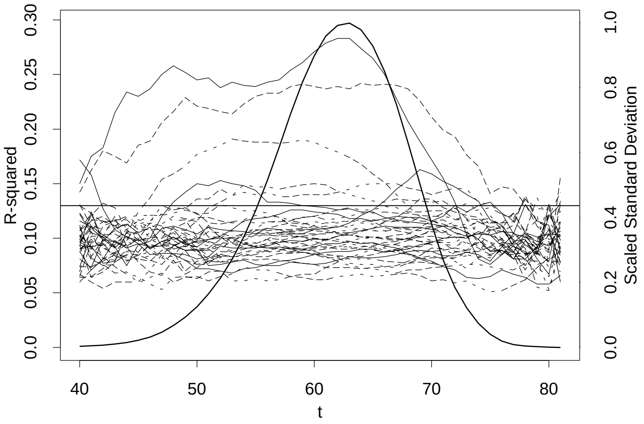


Figure 9.5 The $R^2(t)$ for 36 inputs for CF

Figure 9.5 gives $R^2(t)$ and standard deviation plots for CF. The figure indicates the range of maximum variability for CF—suggested in Figure 9.2—corresponds to t between 50 and 75. Within that range, 3 inputs stand out as important. In all, 8 inputs were identified as important to CF in Stage 1: numbers 30, 31, 33, 35, 47, 48, 59, and 65.

Figure 9.6 shows the similarity between PD and CF. This figure, however, contains an example where an input is indicated as important by R^2 for t between 40 and 50 but not as much so, practically speaking, because of the smaller variability in the values of PD, as indicated by the standard deviation curve. Eight inputs are selected as important for PD: numbers 28, 30, 31, 33, 34, 35, 47, and 48.

In the next stage of screening, previously selected inputs are fixed at their median values and the importance of the remaining inputs is assessed. However, the strategy to follow with three outputs and three different candidate subsets for Stage 1 is not apparent. In fact, different models and analyses require different strategies. The input subsets in Table 9.1 suggest three possible alternatives. First, only the common inputs, 30, 31, and 33, might be fixed. This strategy might work if those inputs were clearly dominant for each output, which is not the case in this application. Second, all 15 inputs might be fixed for the next stage. The problem with that strategy for CF, for example, is that the importance of inputs 38, 40, 42, and others, over and above that of those selected by their R^2-values, will not be known. As a result, the final set of choices for important inputs is likely to be larger than it need be. The final strategy is to proceed with three separate analyses, one for each output. This strategy suffers from the criticism of the second strategy when applied to each value of t and requires a substantial number of computer runs. However, it is a reasonable approach which will provide useful and accurate information.

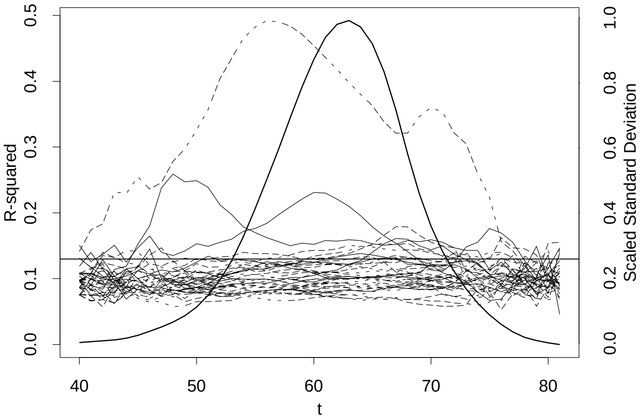


Figure 9.6 The $R^2(t)$ for 36 inputs for PD

9.2.2 Subsequent Stages for EF

When the 10 inputs selected for EF, indicated in Table 9.1, are fixed at their median values, prediction uncertainty is reduced. The first 50 runs for the Stage 2 sample are shown in Figure 9.7. The reduction in variability due to fixing the 10 inputs is apparent by comparing Figure 9.7 with Figure 9.1.

The sample for Stage 2 is analyzed for importance just as that from the first stage except that 10 inputs are at fixed values. The Stage 2 for EF sample of input values is the one from the first stage with the values for the 10 selected inputs replaced by their median values. The R^2-values

Table 9.1 Candidate inputs for EF, CF, and PD

Input #	EF	CF	PD
30	+	+	+
31	+	+	+
33	+	+	+
59	+	+	
65	+	+	
35		+	+
47		+	+
48		+	+
38	+		
40	+		
42	+		
50	+		
27	+		
28			+
34			+

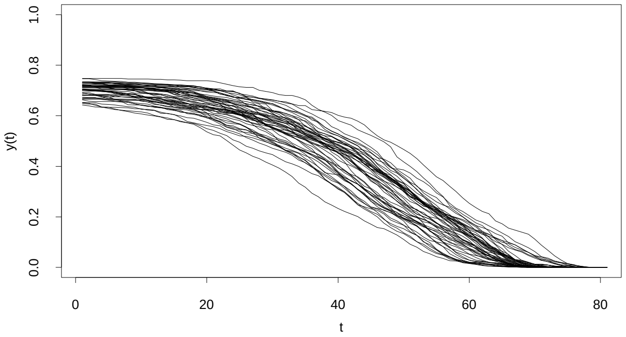


Figure 9.7 Representative $y(t)$ for EF with 10 inputs fixed

computed for the remaining 26 inputs could be estimates of partial correlation ratios except that only one site for the previously selected 10 inputs was used. The R^2-values are properly termed conditional and plotted in Figure 9.8. The scale of the standard deviation is still the maximum from Stage 1. Thus, it is seen that the maximum variability has been reduced by about 40%. Also, there is a single dominant input as indicated by R^2 in a region of lower variability in EF, for $0 < t < 20$. Five additional inputs are indicated in Stage 2: 29, 43, 47, 49, and 62.

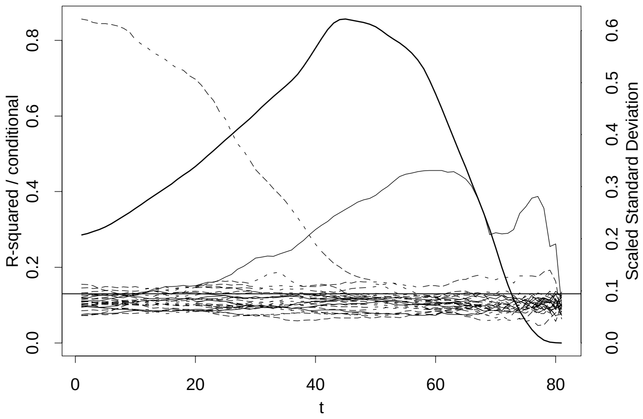


Figure 9.8 Conditional $R^2(t)$ for 26 inputs for EF with 10 inputs fixed

For the next stage, the base case input sample is used but with 15 inputs set to their medians. The first 50 runs for the resulting EF are given in Figure 9.9. The additional reduction in variability, as compared with Figure 9.1, is significant and may be sufficient to terminate the designation of important inputs.

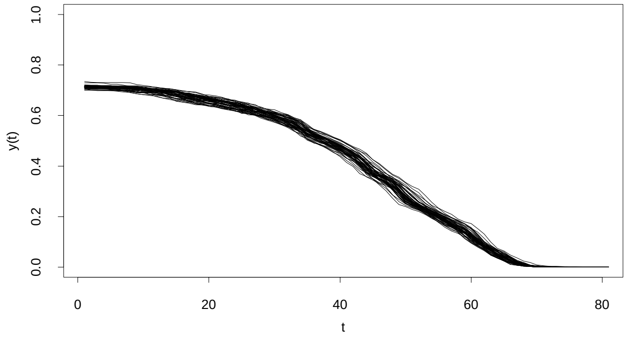


Figure 9.9 Representative $y(t)$ for EF with 15 inputs fixed

Conditional R^2 are computed for the remaining 21 inputs and plotted in Figure 9.10. The scaled standard deviation shows that variability is reduced to 15% of its maximum in Stage 1. The behavior of the plots results from the tight spread in the data and the use of the rank transform. Thus, it is seen that the maximum variability is reduced by about 40%. The additional important inputs indicated in Stage 3 are 25, 45, and 48.

When the selected inputs are fixed at their medians, the total of 18 of the 36 inputs fixed reduces the variability in EF to that given in Figure 9.11. Sequential screening for EF is halted at this stage, to be followed by the validation phase. It is important to remember that as inputs were selected, they were fixed at their median value. Therefore,

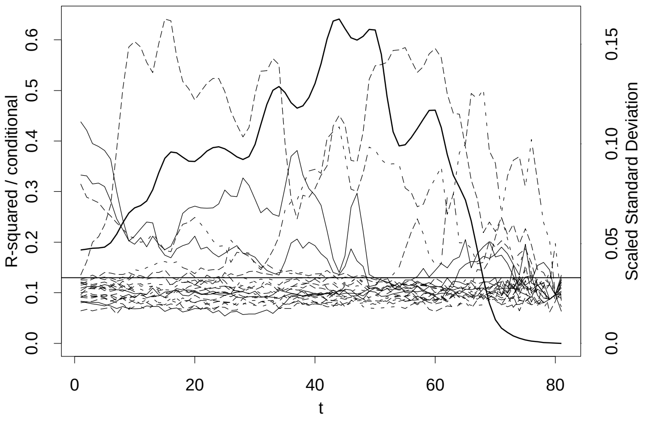


Figure 9.10 Conditional $R^2(t)$ for 21 inputs for EF with 15 inputs fixed

the behavior of the EF for other fixed values is unknown. Additional fixed values (sites) are investigated as part of validation.

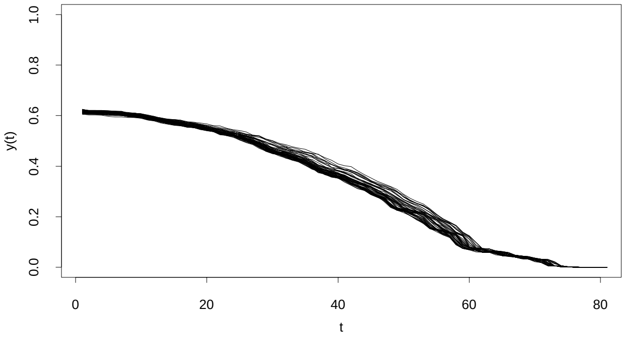


Figure 9.11 Representative $y(t)$ for EF with 18 inputs fixed

9.3 Validation for EF

In the screening portion of the analysis, the set of important inputs for EF was constructed in three stages. The 18 inputs chosen are $S_x = \{25, 27, 29, 30, 31, 33, 38, 40, 42, 43, 45, 47, 48, 49, 50, 59, 62, 65\}$. The objective of the validation portion of the analysis is to determine (1) how much S_x controls variability in EF when S_x is fixed and the other inputs vary, and (2) how much the variability caused by the other inputs obscures the changes in EF caused by S_x, and (3) how well the variability due to S_x mimics the total prediction uncertainty of EF.

Figure 9.12 addresses points (1) and (2). EF from five samples of 50 runs each are plotted in the figure. They are presented as examples of the data that must be investigated for validation and not meant to imply that five is a sufficient number. Unfortunately, no hard and fast rules exist for a sufficient number of validation runs for a general model, so judgement must be exercised.

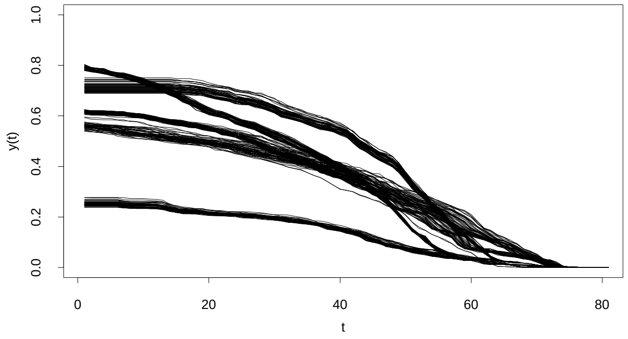


Figure 9.12 Five sets of representative $y(t)$ for EF with important inputs fixed

The spread in each band is due to the inputs of S_x^c. The displacements of the bands is due to the important inputs of S_x. For the application at hand, the figure suggests that the important inputs have been adequately identified. However, in any particular analysis an adequate sample of validation runs must be closely investigated. In this validation procedure, the values of S_x^c are the same 50 for each band; only the values of S_x change from band to band. Thus, intrinsic differences in band patterns are due to interaction between the values of S_x and S_x^c. Investigation of such interactions is often profitable.

9.4 Validation for EF of Selections by Partial Rank Correlation Coefficient

The same sample data used in the preceding analyses can be used to compute partial rank correlation coefficients (PRCCs). Although the PRCC is an established indicator of importance, it relies on assumptions of linearity or monotonicity for it to be effective. The PRCCs are computed for two data sets: the first 50-run replicate in the base case sample and the entire 500-run base case sample. The first 50-run replicate represents a sample size typical of common usage. By using the full 500-run base case sample, the PRCC is given a more even footing with variance ratios. For comparisons, the inputs selected are given in Table 9.2.

Figures 9.13 and 9.14 correspond to Figure 9.11 and are presented for comparison of the relative importance of the inputs selected with PRCCs. The widths of the

Table 9.2 Inputs selected for EF with variance ratios, PRCCs from sample size 500, and PRCCs from sample size 50

Variance	PRCC-500	PRCC-50
25		
27	27	27
29		
30	30	30
31	31	31
33	33	33
38		
40	40	
42	42	42
43	43	
45		
47	47	47
48	48	
49	49	
50	50	
59	59	59
62		
65	65	

bands reflect the importance of the inputs *not* selected and, hence, the adequacy of the selection procedure. Figure 9.13 shows the significant variability not accounted for by the seven inputs selected with PRCCs in the sample of size 50 (PRCC-50). Figure 9.14 shows that the additional six inputs selected with PRCCs in the larger, 500-run sample significantly reduce variability in EF. Whether the remaining variability is significant depends on interpretation. However, a visually substantial amount remains as compared with that in Figure 9.11.

This simple comparative study points out the shortcomings of exclusive use of the PRCC—particularly without validation—as an indicator of importance. However, the PRCC is a valuable adjunct to variance ratios for screening for important inputs. Its use in this manner is recommended. (A simple example demonstrating a complete breakdown of the correlation coefficient to indicate importance is given in Appendix C.)

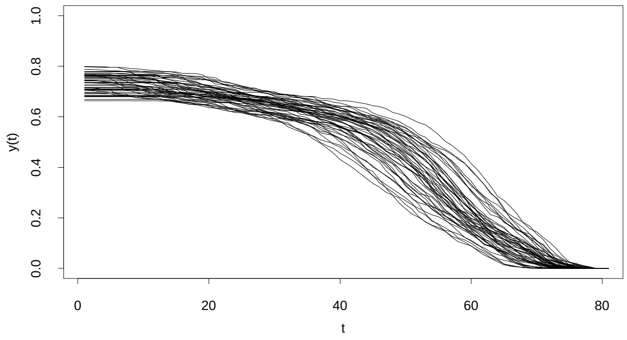


Figure 9.13 Representative $y(t)$ for EF with 7 inputs from PRCC-50 fixed

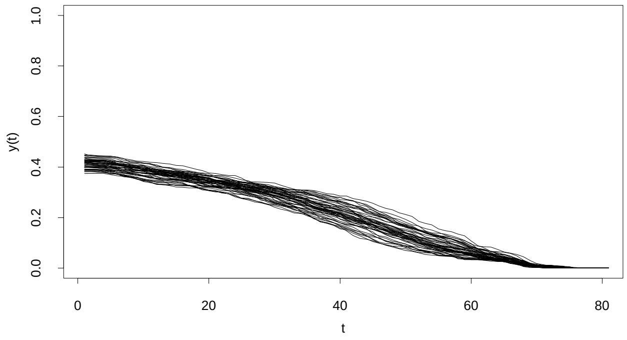


Figure 9.14 Representative $y(t)$ for EF with 13 inputs from PRCC-500 fixed

9.5 Subsequent Stages and Validation for CF

The eight inputs initially selected for CF, indicated in Table 9.1, have values fixed at their medians for the second stage. The first 50 runs for the second stage sample are shown in Figure 9.15.

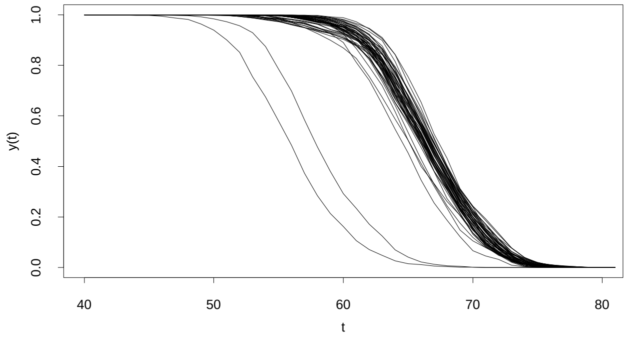


Figure 9.15 Representative $y(t)$ for CF with 8 inputs fixed

The figure displays an interesting pattern with 2 of the 50 runs falling noticeable to the left of the remaining runs. Further investigation of the pattern is not discussed in preference to continuing with variable selection. Conditional R^2 are computed for the remaining 28 inputs and plotted in Figure 9.16. The scaled standard deviation shows that variability is reduced to 60% at its maximum in Stage 1. The six additional important inputs indicated in Stage 2 are 29, 38, 40, 41, 51, and 60.

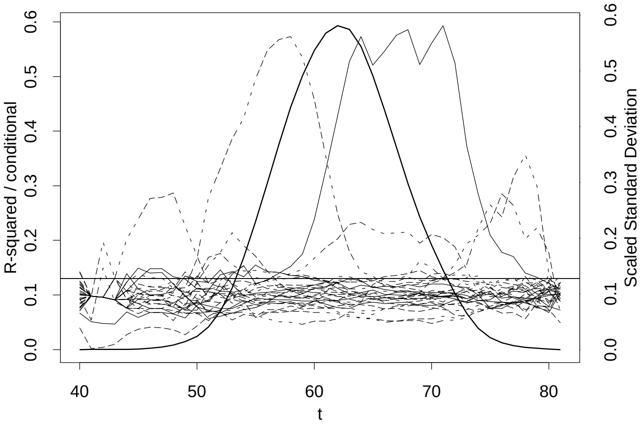


Figure 9.16 Conditional $R^2(t)$ for 28 inputs for CF with 8 inputs fixed

When the 14 inputs selected are fixed at their medians, the variability of CF is very small, as shown in Figure 9.17. The remaining variability is virtually eliminated by the fixing inputs 27, 42, and 50. These last three inputs are noted but not added to the list of important inputs for CF because the additional reduction in variability is not of practical significance.

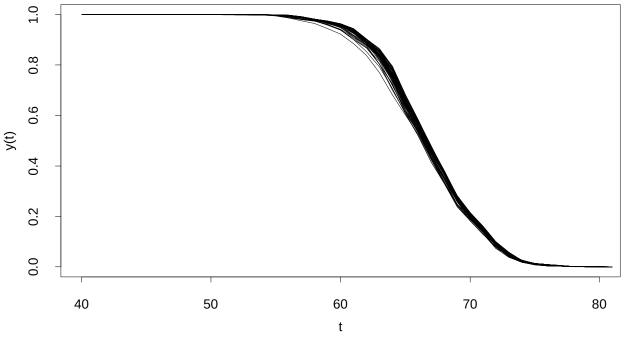


Figure 9.17 Representative $y(t)$ for CF with 14 inputs fixed

The final subset S_x of important inputs for CF contains 14 inputs. For a sample of five sites for S_x, LHS samples of size 50 in the remaining 22 inputs produce the five bands in Figure 9.18. The figure shows how little variability is caused by the 22 unimportant inputs.

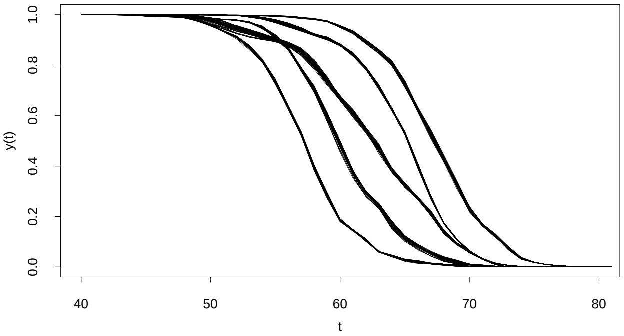


Figure 9.18 Five sets of representative $y(t)$ for CF with important inputs fixed

On the other hand, the variability attributable to the 14 important inputs is expected to be, approximately, the same as the complete prediction uncertainty for CF. A plot similar to Figure 9.18 but with bands corresponding to fixed values of the unimportant inputs S_x^c demonstrates the expectation. Because the bands overlap so completely, band (sample) means and standard deviations are plotted as functions of t in Figure 9.19 as a summary of the bands of 50 sample curves. Ten data sets instead of five are represented in the figure. Together, Figures 9.18 and 9.19 suggest that no important inputs were overlooked in the sequential screening steps.

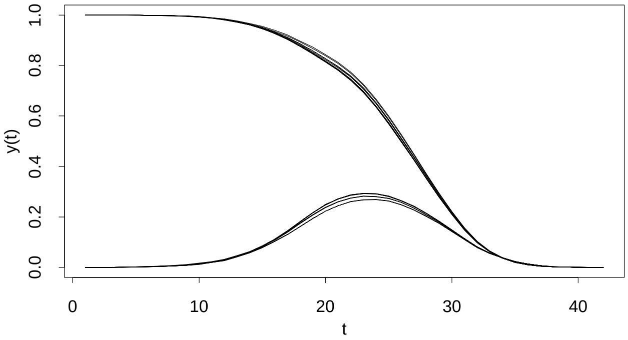


Figure 9.19 Means and standard deviations for CF with important inputs varying

9.6 Comparison of Important Inputs for EF, CF, and PD

Important inputs independently selected in the three stages of sequential screening for each output are given in Table 9.3. Simultaneous screening can mask important inputs. Had screening for the outputs been done in a simultaneous fashion, all inputs indicated by 1 in the table would have been fixed at Stage 2. As a result, the importance of input 47 for EF would have been masked by its selection as important for CF and PD in Stage 1. This example points out a drawback of simultaneous analysis of outputs.

Table 9.3 Important inputs and stages when selected for EF, CF, and PD

Input #	EF	CF	PD
29	2	2	2
30	1	1	1
31	1	1	1
33	1	1	1
38	1	2	2
40	1	2	2
47	2	1	1
48	3	1	1
59	1	1	2
65	1	1	
27	1		2
35		1	1
41		2	2
60		2	2
25	3		
42	1		
43	2		
45	3		
49	2		
50	1		
62	2		
51		2	
28			1
34			1

10 SUBMODEL UNCERTAINTY

This section discusses several techniques to use in evaluating structural uncertainty for submodels. Although easier to approach than general structural uncertainty, the area is in its early stages of development. Therefore, the contributions below address special cases and are to be viewed as tentative. The final topic of the section concerns how one might choose between input and structural uncertainty when comparing model prediction with validation data.

Submodel uncertainty is examined as a special case of structural uncertainty following McKay (1993). A submodel is a meaningful intermediate calculation within the context of the entire model. That is, $s(\cdot)$ is a submodel when it is a function of the inputs and $m(x) = m^*(x, s(x))$ is a nontrivial function of $s(x)$. If $s(\cdot)$ is a function of a subset of the inputs $v \subset x = \{u, v\}$, then the calculation of the output is described as in Figure 10.1. The notation used is that $s(\cdot)$ refers to the structure of the submodel and that $s(v)$ refers to its calculated output value as a function of v.

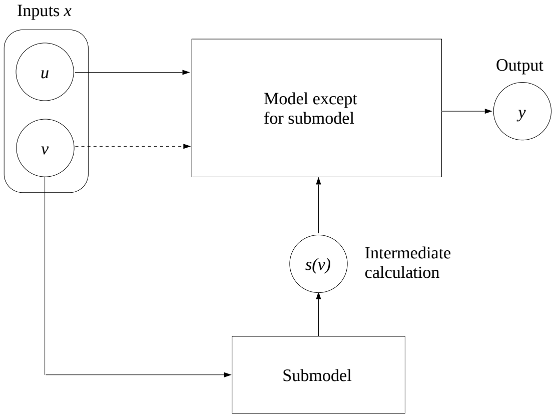


Figure 10.1 Calculation via a submodel

Submodel uncertainty is discussed in two cases. The first supposes there are several known submodels for which relative effects on uncertainty in prediction y are desired. The approach in this case compares the separate analyses where each submodel is used. The second case supposes there are no known alternative submodels and assesses importance of perturbation of the submodel calculation relative to importance of model inputs.

There is an important difference between structural uncertainty in general and submodel uncertainty in the two special cases. Namely, the effect of different submodel calculations—with alternative submodels or perturbations—can be evaluated in changes to the prediction distribution arising from input uncertainty. Consideration is now given to the two cases: the case of competing submodels and the case of perturbation of the calculation of a single submodel.

10.1 Competing Submodels

For the more simple case of competing submodels, an obvious strategy uses the differences within the set of prediction probability distributions f_y from input uncertainty corresponding to each submodel individually. If the probability distribution functions do not differ significantly, submodel choice does not have a substantial impact on prediction uncertainty. The question here is one of practical significance as opposed to statistical significance, although there may be a place for a significance test. In other words, a subjective determination has to be made as to the importance of observed differences among the probability functions relative to the spread of predicted values described by the individual distributions. Specifics regarding methods for comparisons will have to be investigated. Although there do not seem to be any simple, uniquely applicable measures, possibilities include relative entropy (or Kullback-Leibler distance) as discussed by Kullback (1968) and Cover and Thomas (1992) and measures like Hellinger distance and Matusita's distance (see Kotz and Johnson, 1982).

If subjective probabilities are associated with the choices of submodels, two additional options are available for assessing competing submodels. First, the set of prediction probability functions f_y due to the submodels could be viewed as the set of conditional distribution functions from which can be determined the unconditional distribution of y that incorporates submodel uncertainty. This distribution might be used by decision-makers as a summary of prediction uncertainty from the competing models. Second, an indicator input variable which selects from among the competing submodels could be analyzed together with the other inputs as part of input uncertainty. In both options, choice of input distribution function (f_x) might depend on the particular submodel being used. Also, relative advantages and disadvantages of the approaches will depend on the particular instance of model and submodels.

10.2 Perturbation Method for a Single Submodel

In the case of a single submodel, (random) perturbation of its calculation artificially creates the effect of using different submodels. In paralleling the use of a selection indicator variable, a perturbation input variable that varies according to a prescribed probability distribution is used to control the perturbation of the submodel calculation. A perturbed calculation represents, in a sense, an unknown competing model calculation, $\tilde{s}(v)$. Two possible representations for $\tilde{s}(v)$ are an absolute, additive perturbation and, under restrictions, a proportional, multiplicative one as indicated in Equation 10–1.

$$\begin{aligned}\tilde{s}_+(v) &= s(v) + \delta(v) \\ \tilde{s}_\times(v) &= s(v) \times \delta(v) \qquad (10\text{–}1)\end{aligned}$$

Proportional limits like a factor of 2, indicating multiplications by 2 and 1/2, and limits like plus or minus 10% are familiar. So, to simplify discussion and to make computing more convenient, it is assumed that $\tilde{s}(v)$ can be constructed as a fraction of $s(v)$ within prescribed, multiplicative limits. Of course, the assumption fails when $s(v)$ is 0, and it might be better to use the additive form when values of $s(v)$ can be both positive and negative. Whether the additive or multiplicative form is used, key issues are the dimensionality of $s(v)$ and the dependence of δ on v. (When δ is used as a multiplier, it is sometimes referred to as a “dial.”)

10.2.1 Scalar Submodel Output

The first case is that of scalar submodel output, and supposes that absolute limits $L \leq s(v) \leq U$ on its value can be assigned. The perturbed submodel calculation must be restricted to lie within the limits. When a proportional perturbation of the submodel calculation provides adequate variation for the purpose of evaluating submodel uncertainty, a perturbation input variable is used to multiply the submodel calculation. For example, letting the range of the perturbation variable δ be $(1/\lambda, \lambda)$ or $(1-\lambda, 1+\lambda)$ when $\lambda \leq 1$, the submodel calculation $s(v)$ could be replaced by $\tilde{s}(v) = \delta \times s(v)$, where δ is taken to have a uniform distribution on the interval or a loguniform distribution on the interval with logarithmic limits. In either case, limits of variation are defined in terms of λ.

Evaluation of submodel uncertainty for a scalar output can be done for selected values of the range, λ. For any suitable importance measure, the importance of $s(\cdot)$ can be assessed as the (input) importance of the perturbation factor δ for fixed λ. Then, subjective determination of the effect of uncertainty in the submodel can be examined as a function of the range of perturbation λ. In particular, one might identify the value of λ below which uncertainty in the submodel is unimportant relative to input uncertainty.

There is a very special case where inputs v to the submodel calculation are restricted to that calculation, and the model can be written as $m(u, v) = m^*(u, s(v))$. For this case, the dashed line in Figure 10.1 from the v circle is not present. The analysis can be simplified by replacing the submodel calculation $s(v)$ by s_0, an ordinary input variable. The input s_0 would be defined on the interval (L, U). It is not clear what a suitable (sampling) probability distribution for s_0 would be. The actual distribution of $s(v)$ induced by the probability distribution for v might indicate a course of action. Allowing that a distribution can be developed, submodel uncertainty proceeds as a study of s_0 as a part of input uncertainty without using the submodel calculations at all.

10.2.2 Vector Submodel Output

When the submodel has multiple output calculations, $s(v)$ is a vector whose components can be combined with the true inputs and treated as a subset of inputs whose importance is to be assessed. The situation is complicated because it is likely to be unreasonable to let the range of each component of $s(v)$ vary independently to any significant degree. An example of such an output is the wind field calculated by an atmospheric dispersion model. As a first approximation, a bounding box is defined centered at s_0 whose proportions (shape) are determined by a fixed vector δ_s and whose size is determined by a scale parameter λ. The shape of the box defines the allowed proportional variation in the components of $s(v)$. Importance of $s(v)$ is then assessed as a function of the scale, λ, of the box which measures the amount of perturbation.

If there is interaction among components of $s(v)$, the direction of the vector δ_s becomes significant to the analysis. While it may be informative to examine importance as a function of the direction of $s(v)$, high dimensionality of $s(v)$ may make complete analysis difficult. A possible approach is to treat the direction cosines of δ_s as inputs and to perform an analysis in the spirit of principal components.

10.2.3 Nonseparable Inputs

In previous discussions, $s(v)$ or its components could be viewed like model inputs that were independent of the true inputs to the model. The reason was that dependency of $m(\cdot)$ on v was through $s(\cdot)$. Besides examining alternative submodels as arbitrary functions $\widetilde{s}(\cdot)$, they might be examined through and as probability functions defined on the space of $s(v)$, the submodel output. These distributions would be independent of the distribution of the true model inputs. When input to the submodel is nonseparable and appears elsewhere in the model calculation, assessment of effects of submodel perturbation becomes difficult because the probability distribution of submodel output will have to be properly treated as a function of v except in extraordinary situations.

10.3 Components of Error

It is important to be able to decide between input error and structural inadequacy when model prediction and external validation information do not agree. Relevant external information might be available and include experimental data, observational data, or expert opinion.

The first case is for a single model and input distribution. “Best estimate” input values x^* and “best” model $m^*(\cdot)$ predict the data value θ. The best prediction is

$$y^* = m^*(x^*) \ ,$$

for which the absolute value of prediction error is

$$\varepsilon_\theta = |y^* - \theta| \ .$$

In a very formal manner, an allocation process between inputs and model structure might begin with

$$\varepsilon(x, m) = |m(x) - \theta|$$

as the difference for unspecific input x and model $m(\cdot)$. For the model of interest $m^*(\cdot)$,

$$V_x(m^*; \theta) = \{x \mid \varepsilon(m^*, x) \leq \varepsilon_\theta\}$$

denotes the set of input values for which the difference between prediction and data is less that the observed difference ε_θ. The probability content of the set under f_x is denoted by p_x. That is, p_x is the probability content of the set of inputs producing a difference between prediction and data less than ε_θ when f_x is the likelihood for “correct” input values. Similarly, for inputs fixed at x^*,

$$M_m(x^*; \theta) = \{m \mid \varepsilon(m, x^*) \leq \varepsilon_\theta\}$$

denotes the set of models values for which the difference between prediction and data is less that the observed difference ε_θ. The probability content of the set under g_m is denoted by p_m. How one might preceded from this point is the subject of further research.

Two other cases are now presented. They pertain to choosing between two alternative models and choosing between two alternative input distributions. The situations are similar to one of competing models discussed in Section 10.1. In the first case, the objective is to decide which of the two models is more appropriate, relative to data, under the assumption that the correct distribution of input values f_x is known. Figure 10.2 describes the case for a scalar input and output.

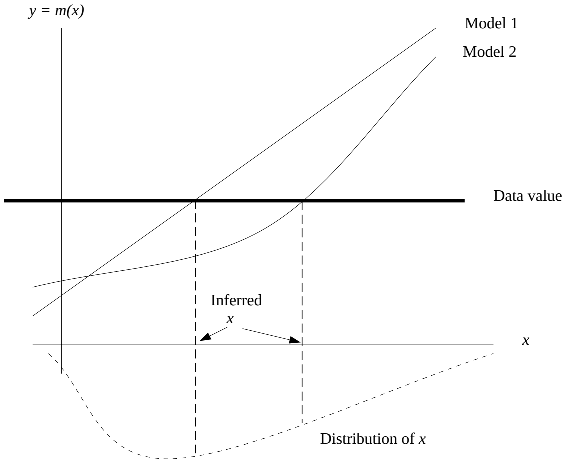


Figure 10.2 Choosing between two models

A strategy to select the model that makes the inferred x-value more likely shows, from the figure, that Model 1 associates an x-value “more likely” relative to the observed data than does Model 2. More likely is in the sense that x for Model 1, indicated by a dashed line from the Model 1 solid line, has a higher likelihood (value of distribution f_x) than the one for Model 2. Thus, based on a likelihood argument, the evidence supports Model 1 over Model 2.

In the second case, the objective is to decide which of the two input distributions is more appropriate under the assumption that the correct model is known. Figure 10.3 describes the situation. In this case, Distribution 1 is selected over Distribution 2 because it assigns a higher likelihood to the inferred x-value that predicts the data.

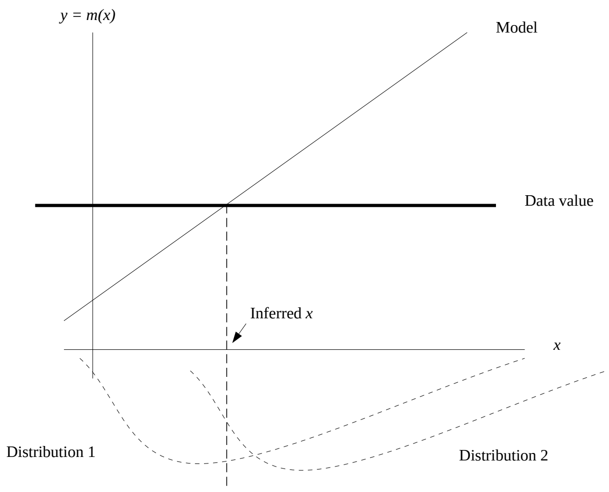


Figure 10.3 Choosing between two distributions

11 CONCLUSIONS

A general mathematical foundation for uncertainty analysis is presented. The foundation provides a reasonable and effective basis to relate prediction uncertainty and importance of inputs through the notion of statistical dependence. As one way of comparing families of conditional prediction distributions, variance ratios arise naturally as importance indicators. Moreover, variance ratios derive their effectiveness directly from consideration of the prediction probability distribution without regard to any specific form of the model $m(\cdot)$. In particular, assumptions of linearity or monotonicity usually accompanying regression-based methods are not necessary.

Although variance is generally preferred over regression-based indicators in evaluation of prediction uncertainty, regression-based methods have served well in many applications. In fact, the auxiliary use of regression-based indicators along with variance-based ones is encouraged.

Estimation for variance-based importance indicators requires special sampling plans. Replicated LHS, as presented in the report, is a viable sampling plan for this purpose. Nevertheless, variance-based methods can require very many computer runs as compared with the number needed for regression-based methods, which require fewer computer runs because of assumptions they make about the form of the model. In cases where variance estimates are unstable because of necessarily small samples, the auxiliary use of regression-based indicators is encouraged.

Importance of individual inputs and subsets of inputs can be determined through sequential screening procedures where importance is indicated with variance ratios, including correlation ratios and partial correlation ratios. The screening process is properly checked through an independent validation step. Validation exercises should be a regular part of an uncertainty study to confirm input selection and to quantify various aspects of prediction uncertainty as related to important inputs.

The report illustrates by way of the analysis applications several important considerations for uncertainty studies. Techniques used in the applications provide useful guidance but will not be applicable in all cases. In general, conditional prediction uncertainty as described by estimated probability density functions or plots of representative values should be examined during input screening to reveal progress of the method and unusual behaviors. Such displays can reveal extreme predictions whose excessive effects on importance indicators is lessened through the use of the rank transformation. When analyzing several model outputs simultaneously in sequential screening, the analyst needs to be alert for possible masking of an input's importance for certain of the outputs due to its selection as important for other outputs.

Finally, the analysis applications pointed out that rank correlation coefficients, both ordinary and partial, can be effective auxiliary indicators of important inputs when used with variance ratios. However, as demonstrated in one of the applications, on their own they can fail to detect important inputs. As a protection, validation of inputs selected as important is effective to confirm input selections, to display the full nature of importance reflected in prediction distributions, and to discover the existence of any undetected important inputs.

12 REFERENCES

Apostolakis, G. (1990). The concept of probability in safety assessments of technological systems. *Science*, 250:1559–1564.

Apostolakis, G. (1993). A commentary on model uncertainty. In *Proceedings of Workshop I in Advanced Topics in Risk and Reliability Analysis, Model Uncertainty: Its Characterization and Quantification*, pages 13–22, Annapolis, MD, October 20–22. U.S. Nuclear Regulatory Commission report NUREG/CP–0138.

Baybutt, P. and Kurth, R. E. (1978). Uncertainty analysis of light-water reactor meltdown accident consequences: Methodology development. Technical report, Report from Battelle's Columbus Laboratories to the U.S. Nuclear Regulatory Commission, Available from the author.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Chanin, D. I., Sprung, J. L., Ritchie, L. T., and Jow, H.-N. (1990). MELCOR accident consequence code system (MACCS): User's guide. Technical Report NUREG/CR-4691, volume 1, U.S. Nuclear Regulatory Commission and Sandia National Laboratories, Albuquerque, NM.

Cover, T. M. and Thomas, J. A. (1992). *Elements of Information Theory*. John Wiley & Sons, New York.

Cox, D. C. (1982). An analytical method for uncertainty analysis of nonlinear output functions, with application to fault-tree analysis. *IEEE Transactions on Reliability*, R-31(5):265–68.

Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26:1–42.

Downing, D. J., Gardner, R. H., and Hoffman, F. O. (1985). An examination of response-surface methodologies for uncertainty analysis in assessment of models. *Technometrics*, 27(2):151–163.

Ford, A., Moore, G. H., and McKay, M. D. (1979). Sensitivity analysis of large computer models — a case study of the coal2 national energy model. Technical Report LA-7772-MS, Los Alamos National Laboratory, Los Alamos, NM.

Helton, J. C., Rollstin, J. A., Sprung, J. L., and Johnson, J. D. (1992). An exploratory sensitivity study with the MACCS reactor accident consequence model. *Reliability Engineering and System Safety*, 36:137–164.

Iman, R. L. and Conover, W. J. (1982). A distribution free approach to inducing rank correlation among input variables. *Communications in Statistics—Simulation and Computation*, B11:311–334.

Iman, R. L. and Helton, J. C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis*, 8(1):71–90.

Iman, R. L., Helton, J. C., and Campbell, J. E. (1981a). An approach to sensitivity analysis of computer models: Part I—introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology*, 13(3):174–183.

Iman, R. L., Helton, J. C., and Campbell, J. E. (1981b). An approach to sensitivity analysis of computer models: Part II—ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology*, 13(4):232–240.

Iman, R. L. and Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, 10(3):401–406.

Karlin, S. and Rinott, Y. (1982). Applications of anova type decompositions of conditional variance statistics including jackknife estimates. *Annals of Statistics*, 10:485–501.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics*, volume 2, chapter 26. MacMillan Publishing Co., New York, fourth edition.

Kotz, S. and Johnson, N., editors (1982). *Encyclopedia of Statistical Sciences*. John Wiley & Sons, New York.

Krzykacz, B. (1990). Samos: A computer program for the derivation of empirical sensitivity measures of results from large computer models. Technical Report GRS-A-1700, Gesellschaft fur Reaktorsicherheit (GRS) mbH, Garching, Republic of Germany.

Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications, New York.

McKay, M. D. (1979). Sensitivity analysis. In *Proceedings of the Workshop on Validation of Computer-based Mathematical Models in Energy Related Research and Development*, Fort Worth, TX. Texas Christian University.

McKay, M. D. (1988). Sensitivity and uncertainty analysis using a statistical sample of input values. In Ronen, Y., editor, *Uncertainty Analysis*, chapter 4, pages 145–186. CRC Press, Boca Raton, FL.

McKay, M. D. (1993). Aspects of modeling uncertainty and prediction. In *Proceedings of Workshop I in Advanced Topics in Risk and Reliability Analysis, Model Uncertainty: Its Characterization and Quantification*, pages 51–64, Annapolis, MD, October 20–22. U.S. Nuclear Regulatory Commission report NUREG/CP–0138.

McKay, M. D. and Beckman, R. J. (1994a). A procedure for assessing uncertainty in models. In *Proceedings of PSAM-II*, San Diego, CA, March 20–25.

McKay, M. D. and Beckman, R. J. (1994b). Using variance to identify important inputs. In *Proceedings of the American Statistical Association Section on Physical and Engineering Sciences*, Toronto, August 14–18.

McKay, M. D. and Bolstad, J. W. (1981). On determining variation and sensitivity in computer models. *Res Mechanica Letters*, 1:171–174.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

McKay, M. D., Conover, W. J., and Whiteman, D. E. (1976). Report on the application of statistical techniques to the analysis of computer codes. Technical Report LA-NUREG-6526-MS, Los Alamos National Laboratory, Los Alamos, NM.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Oblow, E. M. (1978). Sensitivity theory for reactor thermal-hydraulics problems. *Nuclear Science and Engineering*, 68:322–337.

Oblow, E. M., Pin, F. G., and Wright, R. Q. (1986). Sensitivity analysis using computer calculus: A nuclear waste isolation application. *Nuclear Science and Engineering*, 94:46–65.

Parzen, E. (1962). *Stochastic Processes*, page 55. Holden Day, San Francisco.

Pierce, T. H. and Cukier, R. I. (1981). Global nonlinear sensitivity analysis using Walsh functions. *Journal of Computational Physics*, 41:427–443.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435.

Saltelli, A., Andres, T. H., and Homma, T. (1993). Sensitivity analysis of model output: An investigation of new techniques. *Computational Statistics & Data Analysis*, 15:211–238.

Saltelli, A. and Homma, T. (1992). Sensitivity analysis for a model output: Performance of black box techniques on three international benchmark exercises. *Computational Statistics & Data Analysis*, 13:73–94.

Saltelli, A. and Marivoet, J. (1990). Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques. *Reliability Engineering and System Safety*, 28:229–53.

Statistical Sciences (1991). *S-PLUS Reference Manual*. Statistical Sciences, Inc., Seattle, WA.

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2):143–151.

Taguchi, G. (1986). *Introduction to Quality Engineering*. Kraus International Publications, White Plains, NY.

Wong, C. F. and Rabitz, H. (1991). Sensitivity analysis and principal component analysis in free energy calculations. *Journal of Physics and Chemistry*, 95:9628–9630.

Appendix A: ADDITIONAL TECHNICAL CONSIDERATIONS

A.1 A General Variance Decomposition

A variance decomposition used by Cox (1982), attributed to Baybutt and Kurth (1978), consists of a sum of terms depending on subsets of inputs of size 1, 2, and so forth.

$$\begin{aligned}
V[Y] &= \sum_{i}^{m} V_i + \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} \\
&\quad + \cdots + V_{12\cdots m} \\
V_{ijk\cdots} &= V[Z_{ijk\cdots}],\ 1 \le i < j < k < \cdots \le m \\
Z_i &= E[Y \mid X_i],\ i = 1, \cdots, m \\
Z_{kj} &= E\left[Y - \sum_{i=1}^{m} Z_i \mid X_k, X_j\right],\ 1 \le k \le j \le m \\
Z_{kjl} &= E\left[Y - \sum_{i=1}^{m} Z_i - \sum_{n<p} Z_{np} \mid X_k, X_j, X_l\right], \\
&\quad 1 \le k < j < l \le m, \text{ and so forth}
\end{aligned}$$

The first summation in the decomposition is of VCEs. Subsequent terms involve variances of prediction residuals. The expansion looks very promising for importance indication. However, it requires that the inputs be statistically independent. Moreover, there are an excessive number of terms involved, even for a moderate number of inputs.

A.2 Entropy

Variance, information and entropy are related concepts. In particular, relative entropy or Kullback-Leibler distance (Kullback, 1968) could play an important role as an indicator of importance for prediction uncertainty. A helpful discussion of entropy can be found in the *Encyclopedia of Statistical Sciences* (Kotz and Johnson, 1982).

The *entropy* of the density function f_y is defined by

$$\begin{aligned}
H &= -E(\log(f_y)) \\
&= -\int \log(f_y(y)) f_y(y)dy\ .
\end{aligned}$$

The *relative entropy* or *Kullback-Leibler distance* of density $f_{y|s_x}$ relative to f_y is defined by

$$I(s_x) = -\int \log\left(\frac{f_{y|s_x}(y)}{f_y(y)}\right) f_y(y)dy$$

and is a function of s_x. Its expected value is

$$I = -\int \left(\int \log\left(\frac{f_{y|s_x}(y)}{f_y(y)}\right) f_y(y)dy \right) f_{s_x}(s_x)ds_x\ ,$$

which is a measure of the differences among the family of conditional density functions $\{f_{y|s_x}\}$. Thus, I might be used as an importance indicator for S_x.

A.3 Derivation of Equation 5–6 and Motivation for the Partial Correlation Ratio

For the subset $\{x, S_x\}$

$$\begin{aligned}
V[y] &= V_{x,S_x}[E(y \mid \{x, S_x\})] \\
&\quad + E_{x,S_x}(V[y \mid \{x, S_x\}])\ . \qquad \text{(A–1)}
\end{aligned}$$

Conditioned on S_x (e.g., using $f_{y|S_x}$), it is seen that

$$\begin{aligned}
V[y \mid S_x] &= V_{x|S_x}[E(y \mid \{x, S_x\}) \mid S_x] \\
&\quad + E_{x|S_x}(V[y \mid \{x, S_x\}] \mid S_x)\ .
\end{aligned}$$

Expectation over S_x produces

$$\begin{aligned}
E_{S_x}(V[y \mid S_x]) &= E_{S_x}\left(V_{x|S_x}[E(y \mid \{x, S_x\}) \mid S_x]\right) \\
&\quad + E_{x,S_x}(V[y \mid \{x, S_x\}])\ ,
\end{aligned}$$

which gives

$$\begin{aligned}
E_{x,S_x}(V[y \mid \{x, S_x\}]) &= E_{S_x}(V[y \mid S_x]) \\
&\quad - E_{S_x}\left(V_{x|S_x}[E(y \mid \{x, S_x\}) \mid S_x]\right)\ . \qquad \text{(A–2)}
\end{aligned}$$

Substitution from Equation A–2 for the last term in Equation A–1 gives

$$\begin{aligned}
V[y] &= V[E(y \mid \{x, S_x\})] + E(V[y \mid S_x]) \\
&\quad - E_{S_x}\left(V_{x|S_x}[E(y \mid \{x, S_x\}) \mid S_x]\right)\ . \qquad \text{(A–3)}
\end{aligned}$$

Substitution for the second term on the right in Equation A–3 with

$$E(V[y \mid S_x]) = V[y] - V[E(y \mid S_x)]$$

and rearrangement of terms produces Equation 5–6,

$$\begin{aligned}
V[E(y \mid \{x, S_x\})] &= V[E(y \mid S_x)] \\
&\quad + E(V[E(y \mid \{x, S_x\}) \mid S_x])\ .
\end{aligned}$$

A.4 A Useful Derivation Technique

The variance of a sampling distribution can be derived from the expectation of a sum of squares about the sample mean. Sums of squares also appear in analysis tables like the ones in Appendices A.5 and A.6. One way to derive expectations of sums of squares is to use Equation 5–1 as follows. The $y_{j\,k}$ and x_j are defined analogously to those in Appendix A.5. Let

$$\overline{y}_j = \frac{1}{K}\sum_{k=1}^{K} y_{j\,k} \text{ and } \overline{y} = \frac{1}{J}\sum_{j=1}^{J} \overline{y}_j \ .$$

Then,

$$\begin{aligned} E\left(\sum_{j=1}^{J} \left(\overline{y}_j - \overline{y}\right)^2\right) &\simeq JV\left[\overline{y}_j\right] \\ &= J\left\{V\left[E\left(\overline{y}_j \mid x_j\right)\right] + E\left(V\left[\overline{y}_j \mid x_j\right]\right)\right\} \\ &= J\left\{V[E(y \mid x_j)] + \frac{1}{K}E(V[y \mid x_j])\right\} . \end{aligned}$$

A.5 One-way Analysis of Variance Analogy

Let $\{(x_{j,}y_{j\,k}) \mid j = 1, \cdots, n$ and $k = 1, \cdots, r\}$ be sample values with $\{y_{j\,k},\ k = 1, \cdots, r\}$ independent and identically distributed as random variables conditioned on x_j, and the $\{x_j,\ j = 1, \cdots, n\}$ independent and identically distributed. The $\{x_j,\ j = 1, \cdots, n\}$ represent values of a model input x. The $\{y_{j\,k},\ k = 1, \cdots, r\}$ represent the values of the model output for input value x_j with sampled values of the inputs other than x accounted for by the index k. Expected values can be found using the technique of Appendix A.4.

Source of Variation/df	Sum of Squares	Approx. E(Sum of Squares)
Total $nr - 1$	$\text{SST} = \sum_{j=1}^{n}\sum_{k=1}^{r}\left(y_{j\,k} - \overline{y}\right)^2$	$nrV[y]$
Between x $n - 1$	$\text{SSB} = r\sum_{j=1}^{n}\left(\overline{y}_j - \overline{y}\right)^2$	$nrV_x[E(y \mid x_j)] + n\sigma_e^2$
Within x $n(r - 1)$	$\text{SSW} = \sum_{j=1}^{n}\sum_{k=1}^{r}\left(y_{j\,k} - \overline{y}_j\right)^2$	$nrE_x(V[y \mid x]) = nr\sigma_e^2$

$$\begin{aligned} R^2 &= \text{SSB/SST} \\ R_a^2 &= \left(\text{SSB} - \frac{1}{r}\text{SSW}\right)/\text{SST} \\ &= R^2 - \frac{1}{r}\left(1 - R^2\right) \end{aligned}$$

A.6 Two-way Analysis of Variance Analogy

Let $\{(x_t, u_{tj}, y_{tjk}) \mid k = 1, \cdots, r,\ j = 1, \cdots, n,$ and $t = 1, \cdots, s\}$ be sample values with $\{u_{tj},\ j = 1, \cdots, n\}$ independent and identically distributed random variables conditioned on x_t, $\{y_{tjk},\ k = 1, \cdots, r\}$ independent and identically distributed random variables conditioned on x_t and u_{tj}, and $\{x_t,\ t = 1, \cdots, s\}$ independent and identically distributed. The $\{x_t,\ t = 1, \cdots, s\}$ represent values of a model input (subset) x. The $\{u_{tj},\ j = 1, \cdots, n\}$ represent values of another model input u whose probability distribution is conditioned on x. Finally, the $\{y_{tjk},\ k = 1, \cdots, r\}$ represent the values of the model output for input values x_t and u_{tj} with sampled values of the inputs other than x and u accounted for by the index k. Expected values can be found using the technique of Appendix A.4.

Source of Variation/df	Sum of Squares	Approx. E(Sum of Squares)
Total $snr - 1$	$\text{SST} = \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} (y_{tjk} - \overline{y})^2$	$snrV[y]$
Between x $s - 1$	$nr \sum_{t=1}^{s} (\overline{y}_t - \overline{y})^2$	$snrV_x[E(y \mid x_t)] + nr\sigma^2$
Within x $s(nr - 1)$	$\text{SSW} = \sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} (y_{tjk} - \overline{y}_t)^2$	$snrE_x(V[y \mid x_t]) = snr\sigma^2$
Between u within x $s(n - 1)$	$\text{SSB} = r \sum_{t=1}^{s} \sum_{j=1}^{n} \big(\overline{y}_{tj} - \overline{y}_t\big)^2$	$snrE_x\big(V_{u\mid x}[E(y \mid \{x_t, u_{tj}\}) \mid x_t]\big)$ $+ sn\sigma_e^2$
Within u within x $sn(r - 1)$	$\sum_{t=1}^{s} \sum_{j=1}^{n} \sum_{k=1}^{r} \big(y_{tjk} - \overline{y}_{tj}\big)^2$	$snrE_x\big(E_{u\mid x}(V[y \mid \{x_t, u_{tj}\}])\big)$ $= snr\sigma_e^2$

partial $R^2(u; x) =$ SSB/SSW

partial incremental $R^2(u; x) =$ SSB/SST

Appendix B: INPUTS TO MACCS

The names of the MACCS input variables are given below. A few of the names in the table do match those given in the MACCS User's Guide (Chanin et al., 1990) because a newer version of the code was used in this study.

1 NUMFIN	11 RFP1C4	21 RFP2C6	31 ZSCALE
2 TCFMCU	12 RFP1C5	22 RFP2C7	32 VDEPOS1
3 PLUDUR1	13 RFP1C6	23 RFP2C8	33 VDEPOS2
4 PLUDUR2	14 RFP1C7	24 RFP2C9	34 VDEPOS3
5 PLHEAT1	15 RFP1C8	25 BUILDH	35 CWASH1
6 PLHEAT2	16 RFP1C9	26 BUILDW	36 CWASH2
7 PLHITE1	17 RFP2C2	27 SCLCRW	37 TCORUN
8 RFP1C1	18 RFP2C3	28 SCLADP	38 TDELAY
9 RFP1C2	19 RFP2C4	29 SCLEFP	39 LASEVA
10 RFP1C3	20 RFP2C5	30 YSCALE	40 ESPEED
41 P2DOS1	51 EFFACA2	61 TTOSH2	
42 PHS2T2	52 EFFACB2	62 CSFACTS	
43 EVFRAC	53 EFFACA3	63 GSHFACS	
44 CSFACTE	54 EFFACB3	64 PROTINS	
45 CSFACTN	55 EIFACA1	65 EFFTHR1	
46 GSHFACE	56 EIFACB1	66 EFFTHR2	
47 GSHFACN	57 EIFACA2	67 EITHRE1	
48 PROTINN	58 EIFACB2		
49 EFFACA1	59 TIMHOT		
50 EFFACB1	60 DOSHOT		

Appendix C: BREAKDOWN OF THE CORRELATION COEFFICIENT

The correlation coefficient ρ is an effective regression-based indicator of importance under the assumption of an approximately linear relation between model input x and output prediction y. When the actual relation is nonlinear, the correlation coefficient can break down as an indicator of importance. In a simple but realistic example based on an event tree calculation, ρ can fail to detect importance while the correlation ratio η^2 functions correctly to indicate importance. The event tree is indicated in Figure C.1.

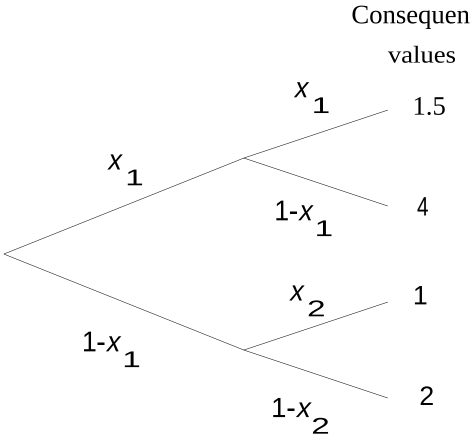


Figure C.1 Simple event tree

Corresponding to the event tree, the model prediction y is the expected consequence which depends on inputs x_1 and x_2. The expected consequence is the sum of the product of probabilities through the tree times consequence values. It is given by

$$y = -2.5x_1^2 + x_1(x_2 + 2) - x_2 + 2 \ .$$

When the inputs x_1 and x_2 have independent uniform probability distributions on the interval $(0, 1)$, the exact values of correlation coefficients and correlation ratios for x_1 and x_2 can be derived to show the following.

- Because y is a nonlinear function of x_1, it is expected that the correlation coefficient may not adequately express the importance of x_1. In fact, this example points out the extreme situation where the covariance between x_1 and y is 0. Therefore, although x_1 is clearly an important input, the correlation between y and x_1 is 0:

$$\rho_1 = 0 \ .$$

- The correlation ratio for x_1 is given by

$$\begin{aligned} \eta_1^2 &= 5/9 \\ &= 2/3 - 1/9 \ . \end{aligned}$$

  The correlation ratio indicates that x_1 accounts for approximately $2/3$ of the (prediction) variance of y.

- The expected consequence y is a linear function of x_2. Therefore, as expected, the correlation ratio and the correlation coefficient squared have to be the same value:

$$\eta_2^2 = \rho_2^2 = 1/3 \ .$$

The example also points out the nonadditivity of the VCE for individual inputs as suggested by the fraction $1/9$ in the value of η_1^2. A PVCE calculation for either input shows the additivity described in Equation 5–11.